SPF Affaires étrangères, Commerce extérieur et Coopération au développement

**Service de l'Évaluation spéciale de la Coopération internationale belge**

# Étude de l'évaluabilité pratique des interventions (co)financées par la Coopération belge

**Nathalie HOLVOET - Liesbeth INBERG - Bob PEETERS - Lisa POPELIER -Dirk VAN ESBROECK -Ellen VERHOFSTADT**

## Rapport final

## Novembre 2015

La présente évaluation a été réalisée par un partenariat composé de South Research et de l'IOB (Université d'Anvers), assistés en cela par un comité d'accompagnement.

Novembre 2015

Conception graphique : Service Communication SPF

Impression : Imprimerie SPF

Évaluation n° S4/2014/03

Dépôt légal : **xxxxxx**

Ce document est également disponible au format PDF en français sur le CD-ROM en annexe et peut aussi être obtenu auprès du Service de l'Évaluation Spéciale ou sur le site web www.diplomatie.belgium.be

Le présent rapport doit être cité comme suit :
Service de l'Évaluation Spéciale / SES (2015), *Étude de l'évaluabilité pratique des interventions (co)financées par la Coopération belge*, SPF Affaires étrangères, Commerce extérieur et Coopération au Développement, Bruxelles.

# Synthèse

## Cadre de l'étude

Cette étude se situe dans le contexte, d'une part, de l'importance croissante des évaluations dans la coopération au développement et d'autre part, du constat selon lequel, pour des raisons diverses, la qualité, l'utilité et l'utilisation effective des évaluations ne répondent toujours pas, à ce jour, aux attentes. Les premières phases de l'étude ont été réalisées par le Service de l'Evaluation Spéciale (SES) lui-même et ont constitué une importante base pour la seconde phase. Le cahier des charges du marché formulait comme suit l'objectif général de l'étude : "*contribuer à rendre toutes les actions futures évaluables à plus ou moins court terme*".

L'étude a eu pour objet, au total, 40 interventions dans 4 pays (Belgique, Bénin, RDC et Rwanda). Il s'agit d'interventions dans différents secteurs, mises en œuvre par un large éventail d'acteurs de développement, avec le (co)financement de l'État fédéral belge. À l'aide d'un cadre d'étude (voir ci-après), un score a été attribué à ces interventions pour une série d'éléments ayant une influence sur l'évaluabilité. Ces scores permettent de comparer entre elles les 40 interventions sur le plan de l'évaluabilité, mais ne permettent pas de tirer des conclusions globales en ce qui concerne le degré d'évaluabilité de ces interventions ou des interventions de la Coopération belge au développement en général.

## Concepts de base

La méconnaissance relative de la notion d'« évaluabilité » et les réserves initiales émises à l'égard de cette étude par certains acteurs ont incité l'équipe d'étude, dans un premier temps, à accorder une grande attention à la création d'une base de soutien pour l'étude et à une description claire des notions « évaluabilité » et « appréciation de l'évaluabilité ». L'étude a utilisé la définition de l'OCDE/CAD, qui définit l'évaluabilité comme étant « *la mesure selon laquelle une activité ou un programme peut être évalué de façon fiable et crédible ».* Quant à l'appréciation de l'évaluabilité, qui est souvent confondue avec l'évaluabilité, elle est décrite comme « *un instrument qui permet de déterminer si une évaluation est indiquée dans une situation donnée* ».

Par ailleurs, une importante distinction est souvent faite entre l'évaluabilité théorique et l'évaluabilité pratique. L'évaluabilité théorique renvoie à l'évaluabilité 'en principe', telle qu'elle peut être déduite de la conception de l'intervention, sans prendre en compte la pratique. L'évaluabilité pratique, quant à elle, prend en compte cette pratique et vérifiera par exemple si les données relatives au progrès d'une intervention sont effectivement collectées et si les systèmes de suivi et d'évaluation (S&E) sont réellement utilisés, de sorte qu'ils puissent apporter une contribution à l'amélioration de la gestion et des résultats.

Le cadre d'étude qui a été développé a tenu compte de ces notions et a été conçu en trois parties, nommées 'dimensions' dans l'étude : le plan d'intervention, la pratique de mise en œuvre et de gestion de l'intervention et le rôle des facteurs contextuels. Ces trois dimensions ont été développées en 8 composants, et ces derniers en 62 items. Ce cadre d'étude constitue un important résultat de l'étude et un instrument utile pour ceux qui souhaitent travailler à l'avenir sur l'« évaluabilité ».

Cette étude a conçu l'évaluabilité avant tout comme un continuum. Selon cette approche, aucune intervention n'est parfaitement évaluable dans tous ses aspects, mais d'un autre côté, il n'existe vraisemblablement aucune intervention qui ne puisse pas du tout être évaluée. En outre, une évaluation peut aussi porter sur différents aspects d'une intervention (les cinq critères classiques de l'OCDE/CAD ou d'autres critères), peut poursuivre différents objectifs (accent mis sur la redevabilité et/ou l'apprentissage, …) et

peut ou non être menée sur la base d'une politique d'évaluation. Dans chacune de ces situations, l'évaluabilité sera différente.

## L'importance de l'« évaluabilité » et de l'« appréciation de l'évaluabilité »

La détermination de l'évaluabilité d'une intervention ne sert *pas* à déterminer la valeur (de développement) de l'intervention en question. Il est parfaitement possible que des interventions de très grande valeur soient difficilement évaluables ; autrement dit, il n'y a aucun lien entre l'évaluabilité (ou le degré d'évaluabilité) et la valeur (de développement) d'une intervention. D'un autre côté, l'importance de l'évaluabilité (la notion) et de l'appréciation de l'évaluabilité (l'instrument) est indéniable, malgré leur application relativement limitée. Leur application peut ainsi permettre d'améliorer la qualité des évaluations et faire en sorte que les résultats des évaluations soient utilisés de manière plus effective. Une bonne appréciation de l'évaluabilité peut être réalisée à un coût qui ne représente qu'une fraction du budget d'évaluation total mais qui, proportionnellement, peut faire une grande différence. En outre – et ceci est aussi confirmé par les résultats de la présente étude – le champ d'application d'une appréciation de l'évaluabilité n'est pas limité à l'évaluation en tant que telle : une telle appréciation peut contribuer à une meilleure gestion dans toutes les phases d'une intervention.

## Principaux résultats et constatations

**Connaissance et utilisation de l'évaluabilité et de l'appréciation de l'évaluabilité.** L'étude a permis de constater que le concept « évaluabilité » et l'instrument « appréciation de l'évaluabilité » (ou « test d'évaluabilité ») étaient, à ce jour, peu connus et utilisés dans la coopération belge au développement. Certains éléments d'un test d'évaluabilité sont mis en pratique ici et là (sans qu'ils soient nommés comme tels), mais nulle part il n'est question d'une application systématique. Cette constatation est significative, car l'« évaluation » a peu à peu gagné en importance ces dernières décennies, jusqu'à faire partie intégrante de la pratique de gestion dans la coopération au développement : les acteurs ne peuvent tout simplement plus se permettre de ne *pas* évaluer. Bien que cette évolution soit globalement positive, elle porte aussi en elle le risque que les évaluations soient ramenées au rang d'exercices rituels, sans implication authentique des acteurs clés. Une utilisation plus délibérée de la notion d'« évaluabilité » et de l'instrument « test d'évaluabilité », avec en corollaire la possibilité d'émettre des jugements fondés quant à l'opportunité d'une évaluation, peut constituer, dans ce contexte, un important instrument pour améliorer – et rendre plus pertinente et réaliste – la manière d'exercer le rôle et la fonction des évaluations dans la coopération au développement.

**Constats globaux.** Les 40 interventions analysées obtiennent un score moyen (sur la base des 62 items du cadre d'étude) qui se situe légèrement au-dessus du point central de l'échelle utilisée. Même si nous ne pouvons pas donner une signification absolue aux scores, ceux-ci constituent une bonne indication de la principale constatation de cette étude, à savoir que les interventions ont dans l'ensemble un certain nombre de points forts, mais aussi un grand nombre de points à travailler si elles veulent améliorer leur évaluabilité. Il ressort également que la distribution des scores est proche d'une distribution gaussienne (normale), mais avec une dispersion importante. Ce qui implique qu'il existe encore, sur le plan de la gestion, de grandes différences parmi les acteurs belges et les types d'interventions, malgré la direction donnée par l'autorité qui assure le financement.

**Les composants forts et faibles et leur influence mutuelle.** Sur les trois dimensions étudiées (plan d'intervention, pratique de mise en œuvre, contexte), la dernière affiche des scores sensiblement plus élevés que les autres. Ceci est une illustration supplémentaire de la marge d'amélioration de ces dimensions, sur lesquelles les acteurs

ont le plus de prise (plan d'intervention et pratique de mise en œuvre). Plus spécifiquement, ce sont les composants 'le système S&E proposé, 'la qualité de l'information de base quant à la mise en œuvre de l'intervention' et 'la qualité de la logique d'intervention et de la théorie de changement' qui présentent les scores les plus faibles. Les scores les plus élevés se retrouvent parmi les composants qui sont repris sous le contexte : 'l'attitude des acteurs clés par rapport à l'évaluation (externe)' et 'l'influence du contexte institutionnel (et politique)'. Une autre constatation est que les imperfections au niveau du plan d'intervention continuent, dans bien des cas, à se faire sentir pendant la mise en œuvre et de cette manière, influencent à la fois directement et indirectement le niveau d'évaluabilité. Une bonne phase de conception est en effet annonciatrice, bien souvent, d'une gestion d'intervention de bonne qualité. Il apparaît que des investissements dans une bonne préparation de l'intervention sont récupérés par la suite. À l'inverse, corriger a posteriori des faiblesses initiales s'avère plus difficile qu'on pourrait le supposer. Autrement dit, le plan initial est la référence sur laquelle s'appuie la pratique.

**Les faiblesses dans le plan d'intervention ont des conséquences différentes pour le suivi et l'évaluation.** Il semble y avoir une différence entre les conséquences d'une phase de conception faible, d'une part pour le suivi, d'autre part pour l'évaluation. À l'évidence, il est possible au niveau – principalement opérationnel – du suivi d'apporter assez facilement des corrections. En témoigne le fait que la qualité du système S&E dans la pratique est sensiblement plus élevée que la qualité du système S&E proposé, même si les lacunes de la phase initiale continuent à se faire sentir dans la mise en œuvre. Les lacunes dans le plan d'intervention s'avèrent toutefois plus conséquentes pour la fonction d'évaluation, parce qu'il est plus difficile d'apporter des corrections. En outre, ces lacunes impliquent que des aspects importants d'une intervention (atteindre effectivement les groupes cibles initiaux, effets de l'intervention sur différents groupes sociaux, réalisation des hypothèses et des risques) ne peuvent (pratiquement) pas être évalués et parfois même, restent totalement en dehors du champ de vision des évaluations. Dans ce cas, la conjonction des facteurs mentionnés ci-dessus peut aussi avoir pour effet que les évaluations 'indépendantes' soient, de facto, fortement dirigées (ou tout au moins déterminées) de l'intérieur. Il en résulte que – volontairement ou non – les vides qui subsistent ne sont pas détectés et les sujets controversés sont écartés ou ne sont pas mis en lumière. En d'autres termes, il y a un risque que les évaluations soient uniquement axées sur la réalité telle qu'elle est définie ou interprétée par l'intervention en question.

**Les scores d'évaluabilité globaux par critère d'évaluation (OCDE/CAD)** donnent, pour toutes les dimensions et tous les composants, des scores plus faibles pour la durabilité et surtout pour l'impact. Le fait que l'on retrouve presque à chaque fois le même schéma dans les scores est évidemment lié aux différents degrés de difficulté – dans toutes les phases du cycle d'intervention – de l'évaluation de ces critères. Il est difficile, sur ce plan, de généraliser, mais on peut affirmer que la durabilité et plus encore l'impact sont plus difficiles à évaluer que les trois autres critères. L'évaluation de l'impact pose des exigences méthodologiques (ainsi que financières) élevées. D'un autre côté, pour l'évaluation de la durabilité, la difficulté est le plus souvent liée au défi que constitue la formulation de jugements fondés par rapport à une situation qui ne se produira que dans le futur. Par ailleurs, il a été constaté qu'en dépit de l'attention croissante accordée à la durabilité, celle-ci est insuffisamment intégrée dans les systèmes de gestion. L'efficacité et surtout l'efficience affichent quant à elles des scores nettement plus élevés, ce qui est une indication de la qualité de la gestion de l'intervention, notamment en ce qui concerne le suivi et l'évaluation. Dans ce cadre, il est important également d'examiner l'influence de l'autorité qui assure le financement : la Direction générale Coopération au développement et Aide humanitaire (DGD) se préoccupe avant tout de l'utilisation correcte des fonds publics mis à disposition. Elle a imposé à cet effet d'importantes conditions (via des procédures, des canevas, …) pour la gestion des interventions financées, ceci en accordant une grande attention aux aspects qui mettent en avant l'efficience, tandis que l'impact, par exemple, reçoit nettement moins d'attention.

Par ailleurs, le bon score pour l'évaluabilité de l'efficience s'explique aussi, certainement, par les efforts importants entrepris notamment par la CTB et les ONG pour développer et mettre en œuvre leurs systèmes de suivi et d'évaluation (S&E). Ces efforts sont la conséquence de processus qui ont cours depuis déjà tout un temps au sein de ces organisations, mais sont aussi une réaction au screening programmé des acteurs non gouvernementaux (ANG), qui aura lieu en 2016 et qui examinera entre autres la qualité des systèmes S&E. Du reste, en ce qui concerne le développement de systèmes S&E, il s'agit dans bien des cas d'initiatives assez complexes qui évoluent de manière graduelle, si bien que les changements au niveau de l'intervention ne sont introduits que progressivement. Dans ces processus, une approche 'bottom-up' (de bas en haut) est suivie, le niveau de la mise en œuvre (inputs – activités – outputs) étant le premier qui entre en ligne de compte ; du reste, c'est aussi le niveau où l'utilité directe pour les organisations concernées est la plus tangible. Il n'est cependant pas évident, pour diverses raisons, que les systèmes S&E intègrent progressivement, de façon "automatique", les niveaux supérieurs de la chaîne moyens-fin et que la fonction d'évaluation, entre autres, soit développée aussi fortement que la fonction de suivi. Car les incitants et les conditions préalables positives qui existent actuellement pour le S&E au niveau de la mise en œuvre (utilité directement démontrable, focus mis traditionnellement sur l'aspect opérationnel, pression de la DGD, exigences relativement moins élevées pour l'organisation et la mise en œuvre) sont moins présents – voire pas du tout – pour le S&E en ce qui concerne les outcomes et les impacts. À politique et contexte inchangés, l'amélioration de l'évaluabilité (concernant en premier lieu les autres critères que l'efficience) n'est absolument pas assurée malgré les progrès enregistrés dans le développement de systèmes S&E ces dernières années.

**Résultats de l'analyse comparative.** Les scores d'évaluabilité par pays présentent très peu de différences entre eux, ce qui implique que d'autres paramètres sont sans doute plus importants. On n'observe – pour des raisons évidentes – des différences notables qu'en ce qui concerne le contexte. Néanmoins, l'influence du contexte sur l'évaluabilité 'technique' n'est pas si grande, dans le sens où elle n'engendre pas d'obstacles majeurs. Du reste, un contexte institutionnel donné peut influer de manière tant positive que négative sur différents aspects de l'évaluabilité, comme l'ont démontré les expériences au Rwanda.
Il n'y a pas de différences fondamentales entre l'évaluabilité des *interventions avec une théorie de changement (TdC) 'complexe' et avec une TdC 'moins complexe'*. Les interventions avec une TdC complexe affichent même des scores légèrement supérieurs, sans doute parce que les acteurs de ces interventions investissent plus dans l'analyse et l'élaboration de systèmes et de pratiques S&E, et parce qu'ils considèrent – à tort ou à raison – que ces interventions sont plus difficiles à financer et que leurs résultats sont plus difficiles à démontrer.
Les scores d'évaluabilité par *canal de financement* (type d'acteur belge) présentent des différences plus marquées que pour la nature des interventions et le pays. On relève en particulier des différences importantes entre 'bilatéral/ONG/syndicats' d'une part, et les 'autres acteurs' d'autre part (meilleurs scores pour le premier groupe), même s'il y a aussi des exemples de bonnes pratiques dans ce dernier groupe. La principale explication de ce constat réside probablement dans les exigences externes (DGD) plus basses (pour les 'autres acteurs') par rapport au plan d'intervention et à la pratique de mise en œuvre. Ceci est renforcé par le fait que pour une partie des acteurs appartenant à ce groupe, la 'coopération au développement' ne constitue pas une tâche principale.

## Principales recommandations

Les différents acteurs impliqués dans la coopération belge au développement ont chacun un intérêt et une responsabilité dans les efforts visant à une meilleure évaluabilité. Ils l'assument, idéalement, au départ d'un cadre et de directives définis et endossés *en commun*, à l'intérieur desquels chaque groupe agit dans son propre rôle et avec sa spécificité. L'équipe d'étude est consciente que certaines des recommandations formulées sont assez ardues, du moins au début. Dès lors, elles ne peuvent être appliquées correctement que si la charge liée à la gestion qui incombe aux acteurs

concernés peut être allégée en conséquence. Ceci se fera de préférence – conformément à la vision de la Note stratégique Résultats de développement – en modifiant le contenu des exigences relatives à la mise en œuvre des propositions d'intervention, des rapports d'intervention, etc., et en portant l'attention sur les résultats de développement (outcomes, impact) plutôt que sur les niveaux opérationnels (moyens, activités, outputs).

Les **recommandations stratégiques** sont les suivantes :

(1) L'intégration systématique, par tous les acteurs, de l'évaluabilite et de l'appréciation de l'évaluabilité, celles-ci étant envisagées comme un moyen pour parvenir à une coopération au développement plus performante et non comme un levier de contrôle ou de direction bureaucratique (par le bailleur de fonds et/ou au sein des organisations). De même, il ne s'agit pas de viser une évaluabilité maximale : l'amélioration de l'évaluabilité doit être une préoccupation permanente, mais elle doit se situer de manière adéquate dans un contexte spécifique ; il y aura toujours un point critique au-delà duquel le coût de la réalisation d'une meilleure évaluabilité ne contrebalance plus les avantages.

(2) L'introduction d'une appréciation de l'évaluabilité cohérente en tant qu'instrument important pour une analyse ex ante de chaque évaluation afin d'analyser et de démontrer les bénéfices potentiels de l'évaluation et d'aboutir, de cette manière, à une décision fondée quant à la réalisation ou non d'une évaluation.

(3) Une amélioration de la phase préparatoire des interventions, en s'attachant plus à la qualité, plutôt que de se limiter à la routine et aux 'vieilles recettes'. Étant donné que ce processus est assez ardu, il est important de mettre en place un changement progressif qui soit soutenu de différentes manières : via un cadre adapté (avec des incitants) de la Direction générale Coopération au développement et Aide humanitaire (DGD), via des bonnes études et évaluations qui pourront soutenir la formulation (p. ex. à la fin des phases précédentes) et via une réduction des exigences administratives et des réglementations (liées aux propositions et rapports d'intervention) qui ne contribuent pas à l'efficacité du développement.

(4) Une attention accrue et une revalorisation des niveaux outcome et impact tout au long du cycle d'intervention (plan d'intervention, Suivi & Évaluation, …) via – entre autres – une définition plus claire de ces notions de base et de la manière dont elles sont concrétisées dans les propositions et rapports d'intervention, et l'élaboration correcte de la théorie du changement avec indication, d'une part, des outcomes directs (intermédiaires) auxquels les interventions peuvent contribuer de façon démontrable sur la base d'une théorie de changement claire et, d'autre part, des effets à plus long terme sur le plan sociétal.

(5) La poursuite du développement des systèmes et pratiques de Suivi & Évaluation (S&E) – qui sont souvent déjà bien élaborés – avec l'ambition d'atteindre une bonne évaluabilité de l'efficacité, de l'impact et de la durabilité. Un tel développement doit idéalement s'opérer de façon graduelle, avec tour à tour une augmentation des moyens, des instruments, de la capacité et de l'expérience, afin que petit à petit, des fonctions plus complexes puissent être reprises et intégrées.

(6) En lien avec le point précédent, il est important qu'un cadre soit créé, dans lequel ces changements (ambitieux) seront non seulement facilités, mais aussi encouragés et valorisés positivement. La DGD joue à cet égard un rôle crucial et pourrait, en accord avec les autres acteurs clés, (a) poursuivre la révision et la simplification de la réglementation, des instruments et des procédures actuels afin qu'ils soient plus axés sur les *effets* (recherchés) de développement, (b) développer des incitants pour pousser plus loin le développement de la fonction S&E (en particulier la fonction d'évaluation), les acteurs étant ainsi mieux en mesure de (faire) réaliser des évaluations de bonne qualité comportant aussi une analyse de la durabilité et de l'impact, (c)

créer un fonds pour le financement d'études et d'évaluations au niveau effet et impact, dont l'initiative émane – de préférence – de l'ensemble des acteurs belges du développement. Ce fonds devrait financer des exercices communs dans lesquels différentes interventions et différents acteurs seraient impliqués et réaliseraient des études et des évaluations qui dépassent les moyens et les capacités des acteurs individuels et/ou qui présentent moins d'intérêt pour eux.

(7) Il convient que la certification planifiée des systèmes S&E ne soit pas dissociée d'une approche plus large et plus intégrée telle qu'elle est par exemple prévue, du moins initialement, pour le screening programmé des acteurs non gouvernementaux (ANG). Plutôt qu'une certification formelle et standardisée, il semble indiqué – conformément à l'approche suivie dans le passé – de développer de bons incitants pour améliorer la qualité de la gestion des interventions et plus spécialement des systèmes S&E. Ces incitants se grefferont de préférence sur les processus déjà en cours parmi les différents acteurs. Un exemple de ces processus est l'élaboration et l'application (par les acteurs eux-mêmes, ou avec un accompagnement externe) d'un instrument de diagnostic permettant aux acteurs concernés (et à la DGD) de se faire une idée des points forts et des points faibles de leur système S&E et de développer un plan par étapes « sur mesure » axé sur le résultat en vue d'améliorer leur système S&E. Pour assurer (durablement) ces fonctions importantes, la DGD doit continuer à disposer des moyens humains (et autres) nécessaires.

L'étude se termine par une série de **recommandations opérationnelles** qui, pour une part, forment une concrétisation des recommandations stratégiques :

(8) Les plans d'intervention doivent accorder plus d'attention à une meilleure description (différenciation) des groupes cibles et à l'élaboration d'une bonne 'baseline'.

(9) La politique S&E et sa traduction dans la pratique doivent être parachevées, avec une plus grande attention pour le développement d'une fonction d'évaluation et pour l'articulation entre la politique S&E de l'organisation même et celle des acteurs locaux.

(10) La fonction S&E au niveau intervention doit encore être développée davantage, ceci en accordant une attention particulière, entre autres, à l'intégration des différentes composantes du S&E dans un système cohérent et à une meilleure implication (en partant du principe de subsidiarité) des acteurs concernés dans le S&E.

(11) Une démarche plus volontaire est indispensable afin d'optimiser l'utilisation finale des évaluations, notamment en impliquant plus étroitement les utilisateurs finaux dans les évaluations (dans toutes les phases) et en planifiant et en réalisant des évaluations selon une approche de type 'portefeuille' qui implique (sur le plan de l'évaluation) une approche différenciée des différentes interventions.

# Table des matières

# Liste des tableaux

# Liste des abréviations

| | |
|---|---|
| APEFE | Association pour la Promotion de l'Education et de la Formation à l'Etranger |
| CA | Comité d'accompagnement |
| CTB | Coopération technique belge |
| CUD | Coopération Universitaire au Développement |
| CAD | Comité d'aide au développement |
| SES | Service de l'Évaluation spéciale |
| RDC | République Démocratique du Congo |
| DFG | Discussion « Focus Group » |
| DGD | Direction Générale Coopération au développement et Aide humanitaire |
| DFID | Department for International Development |
| G&D | Genre et Développement |
| IMEP | Independent Monitoring and Evaluation Project |
| IOB | Instituut voor Ontwikkelingsbeleid en –beheer |
| IMT | Institut de Médecine tropicale |
| S&E | Suivi et Évaluation |
| MIS | Management Information System |
| ANG | Acteur(s) non gouvernemental(aux) |
| ONG | Organisation(s) non gouvernementale(s) |
| OCDE | Organisation de Coopération et de Développement Économiques |
| PCM | Project Cycle Management |
| SR | South Research |
| TdC | Théorie de Changement |
| DTF | Dossier technique et financier |
| UN | United Nations |
| VLIR-UOS | Vlaamse Interuniversitaire Raad, Universitaire Ontwikkelingssamenwerking |
| ONU | Organisation des Nations Unies |
| VVOB | Vlaamse Vereniging voor Ontwikkelingssamenwerking en Technische Bijstand |

# 1. Introduction

## 1.1 Cadre de l'étude

Différentes considérations ont été à la base de l'initiative du Service de l'Évaluation Spéciale (SES) de consacrer une étude à l'« Évaluabilité »[1]. Au cours de la dernière décennie, les évaluations ont pris une place toujours plus importante dans le cycle de gestion, tant au niveau intervention qu'au-delà de ce niveau[2]. Par ailleurs, le rôle des évaluations s'est également clarifié, l'accent étant mis de plus en plus sur les différents objectifs que les évaluations peuvent poursuivre : rendre compte de l'utilisation des moyens reçus, tirer les enseignements des expériences du passé pour faire mieux à l'avenir et soutenir l'élaboration et la mise en œuvre de la politique.

Vu l'importance croissante des évaluations dans la coopération au développement – les évaluations sont considérées comme cruciales, sont souvent obligatoires et ont évolué dans bien des cas vers des systèmes à part entière en lien avec le suivi –, il est très important de s'attacher à la pertinence et à la qualité des évaluations et au suivi de leurs résultats. Une première et importante étape – pourtant souvent oubliée – dans ce contexte consiste à vérifier si toutes les conditions ont été remplies pour exécuter une évaluation de qualité, autrement dit si l'évaluation répond à un besoin existant, si son exécution est réalisable dans le contexte donné et avec les moyens disponibles, et si l'affectation de ces moyens est à la hauteur des bénéfices que l'évaluation est censée générer. D'autre part, il est important de tenir compte du fait qu'il s'est produit une érosion du terme « évaluation » et de sa pratique : bien que l'on ait acquis plus d'expérience en matière d'évaluation, on peut se demander, pour bon nombre d'initiatives qualifiées d'« évaluation », si les exigences minimales de qualité sont bel et bien remplies pour que l'on puisse parler d'une véritable évaluation[3]. Cette érosion est en partie liée à un manque d'expertise mais aussi, et sans doute bien plus encore, au fait que les évaluations font de plus en plus partie du « système » et sont, de ce fait, souvent initiées et mises en œuvre d'une façon (trop) routinière.

Outre ces développements internationaux, les expériences sur le plan belge justifient également que l'on consacre une étude à « l'évaluabilité ». Les expériences en matière d'évaluations externes, mais aussi de suivi, donnent en effet à penser que dans bien des cas, les conditions initiales pour pouvoir bien évaluer ne sont pas présentes, ou ne le sont qu'en partie. On peut dès lors présumer que les exercices d'évaluation qui sont entrepris ne parviennent que très partiellement à atteindre leurs objectifs. D'un autre côté, de nombreuses organisations belges de développement ont investi, ces dernières années, dans le développement de leurs systèmes S&E ce qui, en principe, aurait dû améliorer l'évaluabilité. On dispose cependant de peu de données quant à l'utilité finale de ces systèmes et en particulier des résultats des évaluations. L'« évaluabilité » constitue à cet égard une notion intéressante, car elle peut soutenir le processus décisionnel en ce qui concerne l'opportunité de l'évaluation (d'un projet, d'un programme, …).

---

[1] La notion d'« évaluabilité » est définie plus précisément au chapitre 2.

[2] Dans cette étude, nous utilisons le terme 'intervention' pour désigner l'ensemble des projets, programmes, instruments, … de la coopération belge au développement.

[3] D'un point de vue méthodologique, on attend des évaluations qu'elles soient fiables et valides et qu'elles fournissent des résultats exploitables. Voir aussi le chapitre 2.

Sur la base des constatations qui précèdent, la notion d'« évaluabilité » a gagné en importance ces derniers temps, alors qu'auparavant, elle ne faisait pas l'objet d'une telle attention. Voilà pourquoi le SES a décidé de consacrer une étude à l'« évaluabilité », le service lui-même ayant pris à son compte les deux premières phases de l'étude. Celles-ci comprennent l'élaboration d'un cadre d'analyse de l'évaluabilité théorique et une étude de l'évaluabilité théorique de 43 interventions. Le rapport relatif à ces deux phases a été mis à disposition au début de l'étude (celui portant sur la phase 2 en version préliminaire).

## 1.2 Caractéristiques de base et contexte de l'étude

Avant tout, il est important d'avoir à l'esprit que la nature de ce marché se distingue d'autres marchés lancés par le Service de l'Évaluation spéciale sur plusieurs points importants :

- Ce marché porte sur *une étude et non une évaluation* ; il ne s'agit donc pas, à travers ce marché, de répondre aux objectifs classiques d'une évaluation, mais bien de répondre à *quelques objectifs clairement délimités* (voir 1.3 ci-après) propres à l'étude, formulés avant tout dans une optique d'avenir.
- Comme indiqué précédemment, *les deux premières phases de cette étude ont déjà été exécutées par le SES lui-même*. Elles portent sur l'élaboration d'un cadre d'analyse de l'évaluabilité théorique et sur une étude de l'évaluabilité théorique de 43 interventions.
- Comme l'indique le cahier des charges, *le SES est aussi, exceptionnellement, partie prenante dans cette étude*. Alors que dans un marché SES « traditionnel », le SES est principalement impliqué en tant que fonctionnaire dirigeant, son intérêt dans cette étude est plus grand en ce sens que la réalisation des objectifs de cette étude concerne directement le SES.
- Comme il ressort également de la littérature existante en matière d'évaluabilité, cette notion peut être (et est) utilisée de différentes manières. C'est pourquoi il est important de délimiter clairement le concept d'évaluabilité. Les définitions contenues dans le cahier des charges (ainsi que dans la littérature) concernent aussi bien l'*evaluability* dans le sens restreint du terme que l'*evaluability assessment*, les deux (c.-à-d. la définition et l'instrument) n'étant pas toujours distingués l'un de l'autre de manière stricte. Les définitions mentionnées ont fourni de bons points de référence pour la délimitation de l'étude, mais ont encore dû être clarifiées pendant le déroulement de l'étude (voir le chapitre 2). Dans ce cadre, il ne s'agissait *pas* de mener une discussion exhaustive sur le concept, mais plutôt d'atteindre – de manière pragmatique – un consensus sur la définition de travail à utiliser dans la suite de l'étude.

## 1.3 Objectifs et portée de l'étude

Dans l'approche standard d'une étude ou d'une évaluation, les objectifs et le champ d'investigation sont clairement définis *ex ante* et restent ensuite (quasiment) inchangés au cours de la mise en œuvre. Ce ne fut pas le cas dans cette étude, en ce sens que les objectifs et le champ d'investigation initiaux tels que formulés dans le cahier des charges de l'étude ont été élargis par la suite.

Les **objectifs** tels qu'ils ont été formulés dans le cahier des charges de l'étude (réf. S4/2014/01) peuvent être résumés comme suit[4] :

[4] Voir le point B3, p. 25, du document précité figurant à l'annexe 1. Il est important de signaler à cet égard que les deuxième et troisième objectifs spécifiques – entre autres – ont été largement inspirés par des initiatives qui étaient en chantier lors du lancement de l'étude. Pour diverses raisons, la mise en œuvre de ces initiatives ne s'est toutefois pas déroulée comme prévu initialement.

- L'objectif de l'étude est triple :
  - déterminer, avec les partenaires concernés, les conditions nécessaires, suffisantes et réalistes pour permettre l'évaluation objective d'interventions de coopération ;
  - produire des enseignements utiles à l'harmonisation et la certification des systèmes d'évaluation des acteurs décidées par législateurs ; et
  - vérifier dans quelle mesure les conditions d'évaluabilité mentionnées ci-dessus coïncident (ou non) avec les conditions nécessaires au suivi et à la gestion axée sur les résultats des interventions et avec le cadre légal et réglementaire précité.
- L'objectif général est de (contribuer à) rendre toutes les actions futures évaluables à plus ou moins court terme.

Il est précisé ensuite que les résultats de l'étude doivent, à différents égards, être utiles au SES, aux services de la DG-D et aux Attachés, ainsi qu'aux organismes partenaires de la coopération bilatérale et non-gouvernementale.

Dans la phase initiale de l'étude, il subsistait des incertitudes par rapport aux objectifs précités et, par corollaire, une certaine méfiance parmi une série d'acteurs quant aux intentions finales et aux effets (voulus ou non) de l'étude, exprimée notamment lors de la première réunion du Comité d'accompagnement.

C'est pourquoi l'équipe d'étude a été extrêmement attentive, dans un premier temps, à créer une base de soutien pour l'étude, notamment par une meilleure élaboration du concept d'évaluabilité. En conséquence, le principal objectif de l'étude a été quelque peu rectifié (contribuer à une meilleure évaluabilité des actions de la coopération belge) et l'accent a été mis avant tout sur le potentiel de l'analyse d'évaluabilité pour l'amélioration de la gestion des interventions, en ce compris la détermination de l'utilité d'une évaluation dans une situation spécifique et en général. D'autre part, les objectifs spécifiques susmentionnés ont fait l'objet, dans la mesure du possible et pour autant que ce soit pertinent, de toute l'attention requise.

Enfin, la proposition de l'équipe d'étude de focaliser l'étude pas uniquement sur l'évaluabilité *pratique* comme prévu initialement (c.-à-d. dans le cahier des charges de l'étude), mais à la fois sur l'évaluabilité théorique et pratique, a été acceptée[5].

## 1.4 Structure du présent rapport de synthèse

Le présent rapport de synthèse est structuré comme suit. Ce chapitre introductif est suivi par un chapitre consacré aux concepts d'évaluabilité et d'appréciation de l'évaluabilité. Ce chapitre vise également à démontrer l'importance et l'utilité pratique de ces concepts et l'intérêt qu'il peut y avoir à réaliser un test d'évaluabilité. Ce chapitre s'attache aussi brièvement à la méthodologie de recherche et au cadre d'étude qui est appliqué. Dans les chapitres trois et quatre sont présentés les principaux résultats de l'étude, tandis que le chapitre cinq contient les principales conclusions et recommandations. Le rapport est complété par une série d'annexes : les termes de référence, la description de la méthodologie, le cadre d'étude, la liste des 40 interventions analysées, la bibliographie, la liste des personnes contactées et deux annexes techniques (des explications sur les analyses statistiques et un tableau de cotation étendu).

[5] Ces notions sont précisées au chapitre 2.

# 2. Concepts de base, cadre d'étude et approche

## 2.1 Évaluabilité et appréciation de l'évaluabilité

### 2.1.1 Définitions

La définition de l'évaluabilité la plus utilisée dans la littérature est celle de l'OCDE / CAD (2002 : p. 21) : « *L'évaluabilité est la mesure selon laquelle une activité ou un programme peut être évalué de façon fiable et crédible* »[6]. D'autre part, l'« évaluabilité » et l'« appréciation de l'évaluabilité » sont souvent confondues. Cette dernière notion est alors décrite comme « *un instrument qui permet de déterminer si une évaluation est indiquée dans une situation donnée* »[7]. Cette association indique qu'il est important, non seulement d'utiliser l'évaluabilité en tant que concept, mais aussi de l'associer à la *pratique de l'évaluation* et à la nécessité de vérifier si une évaluation est justifiée et réalisable et est en mesure de fournir des informations utiles. Cela implique que la contribution d'une évaluation à l'amélioration de la gestion et des résultats de l'action soit prise en compte dans l'appréciation de l'évaluabilité.

Avant d'aller plus loin, il est important de disséquer brièvement la définition de l'OCDE/CAD, car la simplicité de cette définition peut être trompeuse, avec le risque de laisser de côté certaines implications importantes. Les notions 'fiable' et 'crédible', qui sont fréquemment utilisées dans le contexte des évaluations mais ne sont pas toujours explicitées (elles ne le sont pas plus, du reste, dans la définition OCDE/CAD), sont ici cruciales :

- Une évaluation est *fiable* lorsque ses résultats de recherche sont stables et cohérents. Cela signifie que si l'étude d'évaluation devait être recommencée, elle conduirait aux mêmes résultats. Les éléments qui influencent la fiabilité sont entre autres la qualité des méthodes d'investigation et de la mise en œuvre de l'étude, la portée de l'étude, l'indépendance des informations et des chercheurs.
- Une évaluation est *crédible* lorsque les résultats de recherche sont considérés comme valides et pertinents par les principales parties concernées (parties prenantes). Les éléments qui influencent la crédibilité sont entre autres : l'attention et la sensibilité pour les dimensions spécifiques au contexte (économiques, culturelles, sociales) de l'intervention ; une large collecte de données avec une triangulation entre les sources et une description détaillée du processus de collecte de données ; la transparence et l'indépendance du processus d'évaluation ; la fiabilité des instruments de mesure ; la consistance et la cohérence dans les résultats et entre les résultats et les conclusions.

Une importante distinction est souvent faite entre '*évaluabilité théorique*' et '*évaluabilité pratique*'. L'évaluabilité théorique renvoie à l'évaluabilité 'en principe', telle qu'elle peut être déduite de la conception de l'intervention, sans prendre en compte la pratique.

---

[6] "Evaluability is the extent to which an activity or project can be evaluated in a reliable and credible fashion"; voir OCDE/CAD (2002) Glossary of Key Terms in Evaluation and Results Based Management. Paris: OCDE/CAD, p. 21.

[7] Voir par exemple : Peter Dahler-Larsen (2013) "Evaluation as a situational or a universal good? Why evaluability assessment for evaluation systems is a good idea, what it might look like in practice, and why it is not fashionable", *Scandinavian Journal of Public Administration*, 16 (3): 29-46.

L'évaluabilité pratique, quant à elle, vérifiera par exemple si les données relatives à l'avancement d'une intervention sont *effectivement* collectées et si les systèmes S&E sont réellement utilisés de sorte qu'une contribution puisse être apportée à l'amélioration de la gestion et des résultats. De cette manière, le lien avec l'utilité (pratique) de l'évaluation est rapidement fait.

Dans cette étude, nous avons abordé l'évaluabilité avant tout comme un **continuum**. Aucune intervention n'est parfaitement évaluable dans tous ses aspects, mais d'un autre côté, il n'existe vraisemblablement aucune intervention qui ne puisse pas du tout être évaluée. En outre, une évaluation peut porter sur différents aspects d'une intervention, à savoir les cinq critères classiques du CAD (pertinence, efficience, efficacité, impact et durabilité), mais aussi d'autres critères. Néanmoins, il n'est pas nécessaire que toutes les évaluations s'attachent simultanément à tous ces critères. Par ailleurs, l'évaluabilité peut aussi être liée au but de l'évaluation. Si les évaluations peuvent avoir différents objectifs (redevabilité, apprentissage, soutien de la politique future), il n'est pas nécessaire qu'ils occupent tous une place centrale dans toutes les évaluations. Si par exemple l'objectif principal est l'apprentissage, les exigences sur le plan de l'évaluabilité ne seront pas les mêmes que si la principale préoccupation est de rendre compte (redevabilité). Enfin, l'évaluabilité dépend aussi de la politique d'évaluation d'une organisation. Dans le cadre de cette politique d'évaluation, on peut par exemple choisir d'utiliser les moyens disponibles pour faire une évaluation stratégique des interventions innovantes, plutôt que d'évaluer toutes les interventions. On peut aussi décider d'investir avant tout dans des évaluations indépendantes centrées sur la redevabilité, plutôt que dans des évaluations internes visant principalement l'apprentissage.

### 2.1.2 Pourquoi l'appréciation de l'évaluabilité est-elle importante ?

Avant toute chose, il est important de préciser que la détermination de l'évaluabilité d'une intervention ne sert **pas** à déterminer la valeur (de développement) de l'intervention en question. Il est parfaitement possible que des interventions de très grande valeur soient difficilement évaluables ; autrement dit, il n'y a aucun lien entre le degré d'évaluabilité et la valeur (de développement) d'une intervention. L'appréciation de l'évaluabilité est plutôt liée à la nature de l'intervention et au contexte dans lequel elle est exécutée. C'est ainsi, par exemple, que l'évaluation d'interventions visant l'autonomisation des femmes comporte généralement des défis plus importants que lorsqu'il s'agit d'interventions portant sur l'approvisionnement en eau potable. Ceci s'explique entre autres par le fait que « l'autonomisation » peut avoir des significations très diverses pour les différents acteurs concernés, qu'il s'agit d'un processus itératif qui comporte plusieurs dimensions et phases, … tandis qu'il est relativement simple de s'accorder sur la signification de « l'approvisionnement en eau ». D'autre part, le contexte joue aussi un rôle important : il est difficile d'effectuer des évaluations dans des zones de conflit ou dans des interventions où les parties concernées sont à couteaux tirés.

Où résident dès lors l'intérêt et l'utilité d'une appréciation de l'évaluabilité ? Bien que le degré d'évaluabilité et la faisabilité d'une appréciation de l'évaluabilité dépendent de différents facteurs (voir plus haut), on peut donner plusieurs réponses générales à cette question :

- Des études relatives à la qualité des évaluations démontrent que celles-ci, dans bien des cas, sont insuffisantes ou génèrent peu de plus-value comparativement à leur coût ; il apparaît en outre que l'utilisation effective des résultats d'évaluation laisse à désirer. Une bonne appréciation de l'évaluabilité peut apporter une solution partielle à ces problèmes.
- Si le but est d'effectuer une évaluation, l'appréciation de l'évaluabilité peut généralement être réalisée à un coût qui ne représente qu'une fraction du budget d'évaluation total ; on peut, de cette manière, éviter que des moyens soient gaspillés, par exemple en évaluations non souhaitables, non réalisables ou mal

conçues. Une telle analyse semble particulièrement indiquée lors de l'exécution d'évaluations complexes, alors qu'elle est moins nécessaire, par exemple, pour des interventions correctement planifiées et mises en œuvre, avec des systèmes S&E qui fonctionnent bien.

- Ensuite, l'intérêt et l'utilité spécifiques d'une appréciation de l'évaluabilité dépendent du moment où elle est exécutée au cours du cycle d'intervention (voir aussi le tableau 1 ci-après) :
    - *Dans la phase préparatoire d'une intervention*, une appréciation de l'évaluabilité portera principalement sur le plan d'intervention : ce plan est-il cohérent et complet, toutes les hypothèses (y compris implicites) ont-elles été prise en compte, concorde-t-il avec l'analyse sous-jacente, … ? Cette appréciation de l'évaluabilité concerne l'évaluabilité *théorique* telle que nous l'avons définie précédemment. Un plan d'intervention de bonne qualité (ou amélioré) sera bien entendu bénéfique à la qualité de l'exécution de l'intervention (y compris le système S&E) et des évaluations ultérieures.
    - *Au lancement d'une intervention* (ou éventuellement juste avant), une appréciation de l'évaluabilité sur la base du plan d'intervention et de la proposition globale d'intervention pourra fournir des éléments précieux pour l'élaboration d'un système S&E.
    - *Pendant l'exécution d'une intervention*, une appréciation de l'évaluabilité peut fournir des indications sur l'opportunité, le timing et la faisabilité d'une évaluation en vérifiant dans quelle mesure les conditions pour l'exécution adéquate d'une évaluation sont remplies (p. ex. quelle est l'attitude des parties concernées ?). Cette appréciation peut aussi examiner si le plan d'intervention est adapté aux développements externes ou aux connaissances acquises et formuler au besoin des suggestions.
    - *Lors de l'achèvement d'une intervention*, l'appréciation sera centrée sur des aspects assez similaires et fournira des indications sur l'opportunité et la faisabilité d'une évaluation (à la différence près que l'on ne pourra plus apporter d'améliorations). L'accent sera mis avant tout sur les défis soulevés par une exécution correcte de l'évaluation, sur la manière de les surmonter et sur les alternatives qui existent à cet égard.
    - *Après une intervention*, les points d'attention de l'appréciation se situent dans le prolongement de ceux qui s'appliquent au moment de l'achèvement d'une intervention ; à cela s'ajoute un aspect important, à savoir la possibilité de contacter les parties prenantes de l'intervention.

**Tableau 1: Utilité et importance d'une analyse de l'évaluabilité**[8]

[8] Basé sur: G. Peersman, I. Guijt, T. Pasanen (2015) *Evaluability Assessment for Impact Evaluation, Guidance, checklists and decision support*. London: Methods Lab Publication, ODI (August 2015).

| Phase du cycle d'iintervention | But de l'analyse | Focus de l'analyse | Résultat envisagé |
|---|---|---|---|
| Formulation de l'intervention | Amélioration du plan d'intervention | Qualité du plan d'intervention | Améliorations et compléments au plan d'intervention |
| Démarrage de l'intervention | Donner un input au développement du système S&E | Disponibilité et qualité de l'information et la collection des données | Améliorations du système S&E (contenu et processus) |
| Mise en œuvre et clôture de l'intervention | • Décider si une évaluation se fera maintenant ou plus tard<br>• Analyser si l'intervention s'est bien adaptée aux évolutions<br>• Donner un input pour le plan d'une évaluation prévue | • Disponibilité, actualité et qualité des données<br>• Position/opinion des parties concernées<br>• Situation dans le contexte plus large | • Compréhension du degré de difficulté de la mise en oeuvre d'une évaluation à ce moment<br>• Formuler des alternatives en ce qui concerne le timing et le contenu<br>• Ajustement du plan d'intervention et du système S&E<br>• Compréhension des choix des objectifs, questions majeures, approche et expertise nécessaire de l'évaluation |
| Après l'intervention | Analyser si la mise en oeuvre d'une évaluation de façon valide et crédible est possible | • Disponibilité et qualité de données en ce qui concerne les effets de l'intervention<br>• Faisabilité de la prise de contact avec des parties concernées | • Compréhension du degré de difficulté de la mise en oeuvre d'une évaluation<br>• Compréhension des possibilités à évaluer (des aspect d') impact et de durabilité<br>• Compréhension du degré de désirabilité d'une évaluation |

Il est clair par ailleurs que l'examen *systématique*, par un organisme de développement, de l'évaluabilité des interventions contribuera à une pratique d'intervention améliorée ainsi qu'à l'élaboration et la mise en œuvre d'une bonne politique d'évaluation, une bonne estimation des interventions à évaluer et des différents objectifs de l'évaluation. Il devient également possible, de cette manière, d'opérer une distinction claire entre « appréciation de l'évaluabilité » et « opportunité d'évaluer ».

## 2.2 Le cadre d'étude proprement dit

Sur la base des objectifs de l'étude, nous avons tenté d'élaborer un cadre d'étude réaliste en tenant compte des moyens disponibles et de la nécessité d'examiner une quarantaine d'interventions. En d'autres termes, cela n'avait guère de sens de proposer un cadre surdimensionné qui, par la suite, n'aurait pas pu être appliqué correctement. Nous voulions éviter, en outre, qu'un cadre trop détaillé conduise à une approche de type « cases à cocher » qui aurait été inévitablement au détriment de la qualité finale et de la pertinence de l'exercice. Concrètement, ceci a conduit au scénario dans lequel, par intervention, un total de quatre jours (visite sur le terrain comprise) était disponible pour déterminer l'évaluabilité, ce qui correspond au nombre maximal de jours qui peut être mis à disposition pour l'analyse de l'évaluabilité d'une intervention[9].

Par ailleurs, nous avons dû, bien évidemment, prendre en considération le fait que « l'évaluabilité » comporte plusieurs dimensions et aspects, comme l'évaluabilité

[9] Soulignons à cet égard qu'une véritable appréciation de l'évaluabilité n'a été que rarement mise en œuvre à ce jour dans le contexte de la coopération belge. Toutefois, la phase préparatoire de nombreuses évaluations contient des éléments qui sont également pris en compte dans une appréciation de l'évaluabilité.

théorique et pratique. Un bon compromis entre 'réalisme' et 'exhaustivité' a été atteint, selon nous, en établissant une hiérarchie dans les différentes composantes du cadre d'étude. Nous avons regroupé, sous une série de notions/critères centraux nommés ci-après '*composants*', les facteurs sous-jacents que nous appelons, dans notre cadre d'étude, les '*items*' (cette approche peut être comparée à celle qui est appliquée dans les évaluations, où les questions principales sont opérationnalisées en critères de jugement, et ces derniers en indicateurs).

Le cadre d'étude permet ensuite d'attribuer un score à tous les composants, ce qui rend possibles l'agrégation et la comparaison entre les interventions, les composantes, les acteurs, les pays, … Le score agrégé peut alors être considéré comme un indicateur de l'évaluabilité, lequel se situe quelque part dans le continuum entre 'évaluabilité nulle' et 'évaluabilité totale'. Lors de l'élaboration du cadre d'étude, nous nous sommes principalement inspirés de l'étude de Rick Davies pour le DFID[10]. En comparaison avec Davies, notre cadre d'étude est cependant plus détaillé. Le cadre se compose de **trois parties** (que nous nommerons ci-après 'dimensions') :

- l'analyse du *plan de l'intervention* (y compris la théorie de changement sous-jacente) par laquelle est prise en compte entre autres (mais pas exclusivement) l'évaluabilité théorique. L'analyse du plan de l'intervention (au sens strict) est également appliquée pour évaluer la qualité du système S&E *proposé* ;
- l'analyse de la *pratique de mise en œuvre et de gestion de l'intervention et du contexte* (y compris l'utilisation des moyens humains et financiers) ;
- l'analyse des facteurs contextuels. Ces facteurs peuvent jouer un rôle à la fois au niveau du plan de l'intervention, sur le plan des systèmes pour la génération des données et lors de la mise en œuvre (future) de l'évaluation proprement dite.

La première partie porte avant tout sur l'évaluabilité théorique (mais pas exclusivement : le système S&E proposé n'en fait pas partie) ; la deuxième partie envisage l'évaluabilité pratique mais ne l'englobe pas entièrement ; la troisième partie, enfin, contient des éléments pouvant influencer l'évaluabilité tant théorique que pratique[11].

Le Tableau 2 donne un aperçu du cadre d'étude et des items auxquels un score a été attribué pour chacun des cinq critères d'évaluation du CAD. Par ailleurs, il est important de noter que pour l'élaboration du cadre, nous avons pris comme point de départ la définition des critères d'évaluation mentionnés qui a été établie par le CAD (voir le cadre).

**Cadre 1 : Les critères d'évaluation du CAD et leur interprétation pour cette étude[12]**

[10] Davies (2013) *Planning evaluability assessments: A synthesis of literature with recommendations* (Report of a study commissioned by the Department for International Development). London: Dfid (October 2013).

[11] Comme indiqué précédemment, le SES a réalisé lui-même une étude de l'évaluabilité théorique destinée à servir d'input pour la présente étude axée sur l'évaluabilité pratique. Étant donné, toutefois, que l'échantillon de la phase d'étude du SES n'était pas tout à fait identique à celui de la présente étude, il est vite apparu qu'il nous faudrait également reprendre l'évaluabilité théorique dans notre étude. C'était la seule manière de faire en sorte que les différents aspects de l'évaluabilité pratique, de même que les liens avec l'évaluabilité théorique, puissent être correctement examinés.

[12] Voir OCDE/CAD (2002) Glossaire des principaux termes relatifs à l'évaluation et la gestion axée sur les résultats. Paris : OCDE/CAD.

**Pertinence** : mesure dans laquelle les objectifs de l'action de développement correspondent aux attentes des bénéficiaires, aux besoins du pays, aux priorités globales, aux politiques des partenaires et des bailleurs de fonds. *L'étude a conservé cette définition mais a été surtout attentive, dans la pratique, aux deux premiers éléments.*

**Efficacité (succès, réussite)** : mesure selon laquelle les objectifs de l'action de développement ont été atteints, ou sont en train de l'être, compte tenu de leur importance relative. *La présente étude a repris cette définition, mais a surtout évalué l'efficacité au niveau des objectifs spécifiques (effet, outcome) des interventions.*

**Efficience** : mesure selon laquelle les ressources (fonds, expertise, temps, etc.) d'une intervention sont converties en résultats[13]. *La présente étude a interprété l'efficience de façon similaire en examinant la manière dont les inputs sont convertis en activités et dont celles-ci débouchent sur des outputs.*

**Impact** : les effets à long terme, positifs et négatifs, primaires et secondaires, induits par une intervention, directement ou non, intentionnellement ou non. *La présente étude est partie de cette définition, mais a aussi, à certains moments, utilisé la notion d'impacts comme étant 'les effets directement attribuables (relation causale) d'une intervention'*[14].

**Durabilité** : continuation des bénéfices résultant d'une intervention après la fin de l'intervention. L'étude est partie de cette définition.

Le cadre d'étude comprend donc au total 3 *dimensions*, subdivisées à leur tour en 9 *composants* et 62 *items* qui, tous, se sont vu attribuer un score[15]. Pour chaque composant et chaque item, un score sur une échelle de cinq a été attribué ; pour les dimensions, la moyenne pondérée des composants a été calculée[16]. Le cadre d'étude complet est repris à l'annexe 3.

**Tableau 2: Présentation sommaire du cadre d'étude**

[13] Dans d'autres documents du CAD, les 'résultats' peuvent être aussi bien des outputs que des outcomes ou des impacts. Dans le compendium, le CAD remplace 'résultats' par le terme plus restreint 'outputs'. Nous avons opté ici pour la définition plus restreinte, car celle-ci permet d'établir une distinction claire entre efficacité et efficience, importante pour cette étude. Il est important d'avoir à l'esprit que ce n'était pas la seule possibilité, et qu'en outre la définition plus large est celle qui est retenue par la plupart des experts en efficience, comme Palenberg qui y a consacré une étude pour le BMZ, fréquemment citée dans le secteur : Palenberg, M., Tools and Methods for Evaluating Efficiency in Development Interventions, BMZ Evaluation Division – Evaluation Working Papers, 2011, 131 p.

[14] Dans la présente étude, la notion d'*impact* est utilisée de deux manières différentes :

en tant que critère d'évaluation au sens strict, ce terme étant distingué (entre autres) de l'*efficacité* et faisant référence aux résultats 'au-delà' du niveau outcome ; dans ce sens, *impact* renvoie à l'objectif général (aux objectifs généraux) tel(s) que formulé(s) dans le cadre logique d'une intervention ;

dans le cadre d'un examen (de la pratique) des '*évaluations* d'impact', en tant que méthode d'évaluation spécifique, l'accent étant mis sur les effets qui sont 'directement attribuables' à une intervention (causalité). Ces effets peuvent se situer aussi bien au niveau outcome qu'aux niveaux plus élevés dans la chaîne. Une évaluation d'impact consiste alors à distinguer les effets qui sont *attribuables* à une intervention (causalité) de ceux qui résultent d'autres facteurs (des facteurs externes ou l'organisation et les instruments de l'évaluation elle-même). Cette interprétation correspond à celle qui est utilisée dans l'évaluation d'impact initiée par le SES : *Évaluer l'impact, la quête du Graal ? Évaluation ex post de l'impact de quatre projets de coopération gouvernementale*. Dans cette étude, l'impact est défini comme suit : les effets qui *découlent* du projet au niveau global. Par ailleurs, les évaluations d'impact qui sont réalisées via le SES sont avant tout des évaluations d'*outcomes*.

Inévitablement, ce double usage de la notion d'*impact* peut être source de confusion ; nous avons tenté, dans la suite de l'étude, de limiter autant que possible cette confusion.

[15] Dans la suite du processus, un de ces neuf composants a été en partie supprimé et en partie intégré dans un autre composant. Le tableau ci-après représente le cadre d'étude adapté.

[16] Voir plus loin ainsi qu'à l'annexe 2 pour plus de détails.

| Critères d'évaluation CAD<br>Dimensions/Composants | Pertinence | Efficacité | Efficience | Impact | Durabilité |
|---|---|---|---|---|---|
| | Nombre d'items | | | | |
| 1. Analyse du plan d'intervention<br>1.1 Analyse sous-jacente (7)<br>1.2 Logique d'intervention et théorie de changement (8)<br>1.3 Le système S&E proposé (9)<br>1.4 La consistance et l'adaptation de la logique d'intervention et la théorie de changement (3) | <br>(7)<br>(1)<br>(5)<br>(3) | <br>(4)<br>(6)<br>(7)<br>(3) | <br>(0)<br>(5)<br>(9)<br>(3) | <br>(0)<br>(5)<br>(7)<br>(3) | <br>(0)<br>(3)<br>(7)<br>(3) |
| 2. Pratique par rapport à la mise en œuvre et la gestion de l'intervention et du context<br>2.1 Information de base quant à la mise en œuvre de l'intervention (11)<br>2.2 Le système S&E dans la pratique (12) | (5)<br>(11) | (10)<br>(12) | (11)<br>(12) | (10)<br>(12) | (8)<br>(12) |
| 3. Le context de l'évaluation<br>3.1 Attidue des acteurs cle (9)<br>3.2 Le contexte plus large (3)<br>3.3 Eléments pratiques (2) (°) | <br>(9)<br>(3) | <br>(9)<br>(3) | <br>(9)<br>(3) | <br>(9)<br>(3) | <br>(9)<br>(3) |
| Scores agrégés | (44) | (54) | (52) | (49) | (45) |
| 4. Suggestion quant au système S&E utilise et à des évaluations futures | | | | | |
| 5. Feedback des concernés sur l'analyse et les suggestion | | | | | |

(°) Cet élément est mentionné pour mémoire uniquement, car ses paramètres et scores sont inévitablement liés à l'organisation d'une évaluation *concrète*.

## 2.3 La démarche de l'étude[17]

### 2.3.1 Échantillonnage

Conformément aux exigences du cahier des charges, 40 interventions dans 4 pays (Belgique, Bénin, RDC et Rwanda) ont été analysées ; dans chaque pays, 10 interventions ont été retenues[18]. Il s'agissait dans tous les cas d'interventions dont la mise en œuvre était déjà bien avancée, éventuellement via une phase antérieure. L'échantillonnage n'a pas été opéré de façon aléatoire, mais en visant une surreprésentation des secteurs et acteurs de moindre importance et des interventions atypiques (p. ex. bourses d'études). Dans l'optique d'une analyse ultérieure, outre le pays, deux paramètres supplémentaires ont été appliqués pour la composition de l'échantillon : interventions complexes par rapport à moins complexes[19] et le canal d'intervention (coopération bilatérale directe, coopération indirecte via ONG et syndicats, coopération via d'autres acteurs).

L'échantillon final se composait comme suit :

- 10 interventions dans chacun des 4 pays ;
- 24 (60%) interventions complexes et 16 (40%) moins complexes ;

[17] L'annexe 2 contient des informations plus précises sur l'approche et la méthodologie de l'étude.

[18] La liste des 40 projets figure à l'annexe 4.

[19] Après un long processus de réflexion interne, le contenu de ce paramètre a été modifié: au lieu d'introduire une distinctions entre des *secteurs* complexes et moins complexes, nous avons opté pour une distinction entre des interventions avec une TdC complexe et moins complexe. Si, par exemple, une intervention dans un secteur moins complexe avait une TdC complexe, cette intervention a été considérée comme complexe. Il est aussi important de remarquer que cette distinction diffère de la distinction entre des interventions complexes et moins complexes *en tant que telles.* Dans ce cas-là, d'autres facteurs interviennent également, comme la taille de l'intervention, le nombre d'acteurs concernés, etc.

- 10 (25%) interventions bilatérales, 20 (50%) interventions exécutées par des ONG ou des syndicats et 10 (25%) interventions réalisées par d'autres acteurs (APEFE/VVOB, universités, BIO, …).

### 2.3.2 Collecte et analyse des données

La collecte des données a consisté à collecter une combinaison de données secondaires et primaires. En pratique, les principales méthodes utilisées sont les suivantes :

- étude et analyse des documents de base (la proposition d'intervention, la baseline éventuelle, les rapports de mise en œuvre et les évaluations, les documents relatifs au système S&E) ;
- interview avec les acteurs clés impliqués dans la mise en œuvre des interventions ;
- discussions en 'groupes focus' (au niveau acteurs et au niveau intervention).

Sur la base des données collectées, le cadre d'analyse de chaque intervention a été complété par l'attribution d'un score aux composants et aux items. Afin de limiter la subjectivité, chaque intervention (à l'exception de 5 interventions en Belgique) a été cotée par au moins deux évaluateurs et en cas de scores divergents, un consensus a été recherché par la discussion. Les scores ont ensuite été rassemblés dans des tableaux, sur lesquels ont été réalisées une série d'analyses statistiques simples et plus complexes (voir les annexes 2 et 7), ce qui a fourni les bases pour les résultats présentés dans les deux chapitres suivants.

### 2.3.3 Aperçu des différentes phases de l'étude

La durée totale de l'étude, jusqu'à la finalisation du rapport final, a été d'environ huit mois (de fin février à mi-octobre 2015). La phase de démarrage de l'étude a pris un temps relativement long en raison du caractère assez atypique du sujet de l'étude et des résistances manifestées initialement par certains acteurs (voir plus haut). Proportionnellement, beaucoup de temps a été consacré à la présentation des concepts de base et de l'approche de l'étude, notamment via l'élaboration d'une note explicative. L'étude des documents de base s'est déroulée dans la période de mars à mai, avec une visite-pilote sur le terrain au Rwanda en avril-mai et des visites au Bénin et en RDC en mai-juin. Les projets en Belgique ont été analysés principalement en juin et juillet. La rédaction du rapport de synthèse provisoire s'est déroulée essentiellement en juillet et août et le rapport final provisoire a ensuite été examiné fin septembre au sein du comité d'accompagnement.

Des réunions se sont tenues avec un comité d'accompagnement à plusieurs moments charnières au cours du processus : au lancement de l'étude, après la première visite de terrain et après la finalisation du rapport de synthèse provisoire.

# 3. Principaux résultats globaux

## Vue d'ensemble

Dans ce chapitre, nous examinons les résultats dans leur *globalité*, c.-à-d. en ce qui concerne les 40 interventions *dans leur ensemble*. Cette analyse globale conduit à des conclusions globales que nous devrons, dans certains cas, nuancer par la suite (chapitre 4) en fonction du pays concerné, du canal d'intervention ou du niveau de complexité (de la TdC) de l'intervention.

Le tableau ci-dessous (tableau 3) donne un aperçu des scores moyens et de l'indice d'évaluabilité en ce qui concerne les huit composants et les cinq critères d'évaluation qui ont été présentés brièvement ci-avant (voir 2.2). Comme indiqué précédemment, chaque composant a été opérationnalisé en une série d'items (le nombre est mentionné entre crochets) tandis que les scores indiqués par composant ne représentent *pas* la moyenne des scores des items par composant, mais bien les scores attribués par les chercheurs au niveau de chaque composant, et ce pour les 40 interventions.

Pour chaque composant (et chaque item), un score de 1 à 5 a été attribué, la valeur 1 correspondant à une mauvaise prestation et la valeur 5 à une bonne prestation. Pour le premier composant (*L'analyse sous-jacente*), on attribuait par exemple un score de 1 si une telle analyse n'avait pas été exécutée/rédigée, tandis qu'un score de 5 était attribué si cette analyse était correcte et complète et soutenait les objectifs de l'intervention. Un score de 3 signifie que l'analyse était bien développée, mais était incomplète (p. ex. pas d'analyse des relations de genre). Dans certains cas (les cellules hachurées), aucun score n'a été attribué parce qu'il n'y avait pas de lien (ou seulement un lien très indirect) entre le composant et le critère d'évaluation. Les résultats plus spécifiques au niveau composant et au niveau item sont présentés en détail et analysés plus loin dans ce chapitre.

Dans ce tableau, et dans d'autres tableaux similaires dans la suite de ce rapport, nous utilisons des couleurs pour marquer plus clairement les différences dans les scores. Ces couleurs permettent de se faire, *de manière très globale, une première idée* de l'évaluabilité relative des interventions. Plus le score ou l'indice est élevé, plus l'évaluabilité est grande. Il est important, à cet égard, de garder à l'esprit que la valeur de l'indice ne doit pas être envisagée en *termes absolus* : elle ne peut s'utiliser que dans un sens *relatif*, à savoir pour comparer *entre eux* les différents critères et composants et pour comparer *entre elles* les 40 interventions. Ainsi, nous pouvons constater que sur l'ensemble des composants, '*le contexte plus large*' obtient le meilleur score tandis que '*le système S&E proposé*' affiche le score le plus faible. Nous pouvons voir également que 'l'impact' est le critère avec le score le plus faible (pour presque tous les composants). Le tableau permet aussi de découvrir au premier coup d'œil le score le plus faible : le système S&E proposé au niveau impact.

### Tableau 3: Index d'évaluabilité par critère CAD et par composant pour les 40 interventions[20]

| | Pertinence | Efficacité | Efficience | Impact | Durabilité | Index d'évaluabilité |
|---|---|---|---|---|---|---|
| **Dimension 1 (plan d'intervention) (°)** | **3,11** | **3,22** | **3,29** | **2,52** | **3,15** | **3,16** |
| 1.1 L'analyse sous-jacente (7) | 3,65 | 3,60 | | | | 3,63 |
| 1.2 La logique d'intervention et la théorie de changement (8) | | 3,00 | 3,50 | 2,10 | 3,25 | 2,96 |
| 1.3 Le système S&E proposé (9) | 2,45 | 2,83 | 2,85 | 1,95 | 2,46 | 2,51 |
| 1.4 Consistance et adaptation de la logique d'intervention et la théorie de changement (3) | 3,55 | 3,44 | 3,52 | 3,50 | 3,75 | 3,54 |
| **Dimension 2 (la pratique de mise en oeuvre)** | **2,99** | **3,25** | **3,51** | **2,33** | **2,73** | **2,96** |
| 2.1 Information de base quant à la mise en oeuvre de l'intervention (11) | 2,95 | 3,20 | 3,53 | 2,13 | 2,62 | 2,88 |
| 2.2 Le système S&E dans la pratique (12) | 3,03 | 3,30 | 3,50 | 2,53 | 2,85 | 3,04 |
| **Dimension 3 (contexte)** | **3,89** | **3,98** | **3,98** | **3,84** | **3,89** | **3,91** |
| 3.1 L'attitude des acteurs clés (9) | 3,68 | 3,70 | 3,70 | 3,58 | 3,63 | 3,66 |
| 3.2 Le contexte plus large (3) | 4,10 | 4,25 | 4,25 | 4,10 | 4,15 | 4,17 |
| **Sc Score global évaluabilité ore Score global évaluabilité (°°)** | **3,26** | **3,38** | **3,52** | **2,70** | **3,13** | **3,23** |

(°) Les scores concernant l'évaluabilité au niveau dimension sont des moyennes non pondérées des scores au niveau composants.
(°°) Pour le score global d'évaluabilité, une moyenne pondérée a été calculée, avec un poids identique pour la dimension 1 et la dimension 2, tandis que la dimension 3 compte pour la moitié de chacune des autres dimensions.

| |
|---|
| Index 4,01 – 5,00 |
| Index 3,01 – 4,00 |
| Index 2,01 – 3,00 |
| Index 1,01 – 2,00 |

Le tableau ci-dessus nous permet de formuler une série de **constats globaux** en ce qui concerne les différentes dimensions et les différents composants du cadre d'étude et les critères d'évaluation :

- On remarque d'emblée qu'à l'exception de la dimension 3 (le contexte), aucune dimension, composant ou critère n'obtient un score vraiment bon : les scores les plus élevés se situent aux alentours de 3,50 et bon nombre de scores se situent dans la 'zone jaune', c.-à-d. sous le point central de notre échelle de valeurs. Cela montre qu'il reste encore, globalement, une marge d'amélioration substantielle.
- Sur le plan des trois **dimensions**, nous pouvons constater que la dimension 3 (le contexte) affiche des scores sensiblement plus élevés que les deux autres dimensions, et que parmi ces deux dernières, la pratique de mise en œuvre (dimension 2) enregistre des scores inférieurs au plan de projet (dimension 1)[21].

[20] Nous n'avons pas toujours été capables de donner un score pour chaque composant et critère CAD. Le tableau A1.2 de l'annexe 7 indique quand et dans quelle mesure une intervention a été scorée.

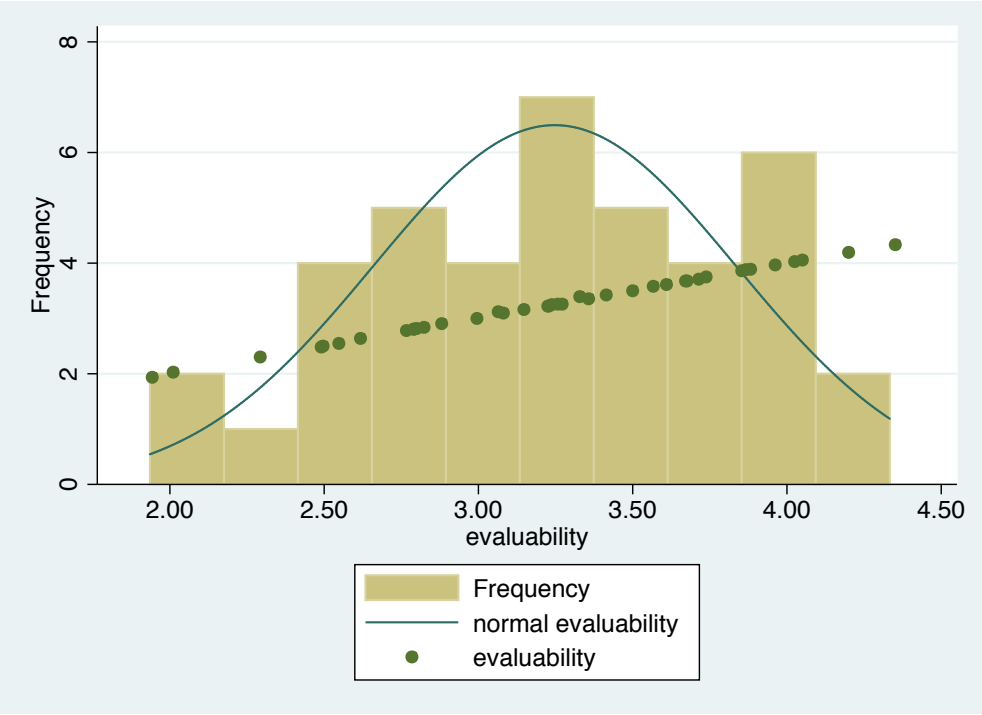
[21] Nous formulons cependant, plus loin (voir la partie 3 de ce chapitre), d'importantes remarques par rapport au bon score des composants et des items sous le 'Contexte'.

- Sur le plan des **composants**, nous relevons à nouveau le bon score des deux composants qui font partie du contexte. Les composants les plus faibles sont le système S&E proposé (dimension 1) et l'information de base quant à la mise en œuvre de l'intervention (dimension 2).
- En ce qui concerne les **critères d'évaluation**, le score global d'évaluabilité le plus faible est celui du critère 'impact'. Les scores sont aussi relativement faibles pour la durabilité et sont relativement élevés pour l'efficacité et surtout pour l'efficience. En réalité, les scores presque identiques – pour tous les critères d'évaluation – de la troisième dimension (contexte) font en sorte que les différences entre les cinq critères, en particulier entre l'impact d'une part et les quatre autres critères d'autre part, sont globalement un peu atténuées. En d'autres termes, si l'on fait abstraction de la dimension 3 (le contexte), les différences entre les critères d'évaluation sont plus marquées. Voici quelques autres constatations :
    - L'évaluabilité de l'*efficience* affiche des scores relativement bons avec, pour presque tous les composants, un score satisfaisant (vert). Le seul composant plus faible est ici le système S&E proposé (dans le plan de projet), mais ceci est compensé par un assez bon score pour le système S&E dans la pratique. Le score relativement bon pour l'efficience est une illustration de ce que nous décrirons par la suite comme une relativement bonne pratique S&E au niveau opérationnel.
    - En ce qui concerne l'évaluabilité de l'*efficacité*, la plupart des scores se situent, ici aussi, dans la zone des scores 'satisfaisants' (vert clair), à l'exception du 1.2 et du 1.3 qui s'y rattache (logique d'intervention et théorie de changement ; système S&E proposé). Cette relative faiblesse révèle une attention trop limitée – entre autres – pour le niveau outcome dans le plan de projet.
    - En ce qui concerne la *durabilité*, les faibles scores pour le système S&E (1.3 et 2.2) révèlent un manque d'attention pour la durabilité dans le système S&E.
    - En ce qui concerne le critère *impact*, le tableau montre que l'évaluabilité de l'impact obtient des scores très modérés à faibles pour tous les composants qui peuvent être liés à la gestion de l'intervention (dimensions 1 et 2)[22]. Pour ces faibles scores, nous identifierons ci-après diverses causes reliées entre elles, essentiellement liées à l'attention réduite qui est accordée au niveau impact par toutes les parties concernées. Le faible score d'évaluabilité pour l'impact est évidemment conforme à ce que l'on pouvait attendre compte tenu des défis méthodologiques de l'évaluation de l'impact, mais aussi – ce qui est plus important dans le cadre de cette étude – de l'attention relativement réduite accordée par les acteurs au niveau impact.

Enfin, nous avons aussi vérifié globalement ce qu'il en est de la **distribution** des scores d'évaluabilité globaux des 40 interventions étudiées (voir la figure ci-dessous). La figure montre une répartition (statistiquement) plus ou moins normale, la fréquence d'interventions la plus élevée se situant aux alentours du score moyen (3,23) de l'indice d'évaluabilité. Un autre élément marquant est l'étalement relativement grand (score minimum de 1,94 et score maximum de 4,33), ainsi que le nombre relativement grand d'interventions (20% de l'échantillon) avec un score élevé (4,00 ou plus).

[22] Nous faisons ici abstraction du composant 1.4, où le nombre d'interventions prises en considération (10 au total) est trop réduit pour pouvoir y associer des conclusions.

**Figure 1 : Distribution des scores d'évaluabilité entre les 40 interventions**




## 3.1 Analyse du plan d'intervention

Dans cette partie sont passés en revue successivement : l'analyse sous-jacente (7 items), la logique d'intervention et la théorie de changement (8 items), le système S&E proposé (9 items) et la consistance et l'adaptation de la logique d'intervention et de la théorie de changement (3 items).

### 3.1.1 L'analyse sous-jacente

La qualité de l'analyse sous-jacente a une influence sur la qualité des évaluations qui portent sur la pertinence et l'efficacité : sans une bonne analyse, en ce compris une analyse (de la situation) des groupes cibles, une évaluation permet difficilement d'établir si une intervention spécifique est pertinente pour les groupes cibles et a effectivement l'effet visé sur les groupes cibles. Cependant, le lien (éventuel) entre l'analyse sous-jacente et l'évaluabilité des autres critères CAD (efficience, impact et durabilité) est moins évident. C'est pourquoi, pour cette composante, contrairement aux autres, des scores ont été attribués uniquement pour la pertinence et l'efficacité.

Sous cette composante sont examinés sept items qui sont repris dans le tableau 4 ci-dessous avec leur score.

### Tableau 4: Résultats principaux quant à l'analyse sous-jacente

| | Pertinence | Efficacité | Efficience | Impact | Durabilité | Index d'évaluabilité |
|---|---|---|---|---|---|---|
| 1.1 L'analyse sous-jacente | 3,65 | 3,60 | | | | 3,63 |
| 1.1.1 Les groupes cibles sont clairement délimités et décrits | 3,48 | 3,50 | | | | 3,49 |
| 1.1.2. Le bien fondé de l'intervention et la situation (problématique) de l'intervention sont clairement décrits | 3,85 | | | | | 3,85 |
| 1.1.3. Le rôle du (des) groupe(s) cible(s) est clairement décrit | 3,60 | 3,55 | | | | 3,58 |
| 1.1.4. Le rôle des acteurs concernés majeurs (en dehors du groupe cible) est clairement décrit | 3,53 | 3,47 | | | | 3,50 |
| 1.1.5. L'analyse genre fait partie intégrante de l'analyse sous-jacente | 2,40 | 2,40 | | | | 2,40 |
| Le lien entre l'analyse sous-jacente et les objectifs de l'intervention est clairement décrit | 4,00 | | | | | 4,00 |
| 1.1.7. Le lien entre le bien fondé de l'intervention et la politique sectorielle du pays partenaire est clairement décrit | 4,10 | | | | | 4,10 |

***Remarque importante** : une version plus élaborée du tableau ci-dessus (et des autres tableaux similaires qui suivent) figure à l'annexe 8 avec, pour chaque score, des explications sur le plan du contenu. Ces explications fournissent une justification du score et sont basées sur un protocole d'appréciation qui a été utilisé par les chercheurs et qui devait leur permettre d'attribuer leurs scores d'une façon cohérente et plus objective. Le lecteur intéressé par le contenu factuel de chaque score est donc invité à consulter l'annexe 8.*

Comparativement aux sept autres composants, *l'analyse sous-jacente* affiche des scores relativement bons, les autres composants de la partie 1 et de la partie 2 présentant des scores moins élevés. Seuls les deux composants sous le contexte de l'évaluation (partie 3) obtiennent des scores supérieurs (voir Tableau 3). Dans de nombreux cas, l'approche est le résultat d'une pratique qui s'est développée et a pris forme au fil des ans, avec une répartition des tâches entre l'organisation belge, (éventuellement) sa représentation locale et les acteurs locaux (voir Cadre 1).

*Cadre 1 : Analyse sous-jacente et développement d'un système S&E chez Vredeseilanden Benin*

Vredeseilanden a pris la décision stratégique de se spécialiser dans l'appui à quelques filières agricoles bien spécifiques. Afin de garantir la cohérence des actions sur le terrain, l'organisation a élaboré des procédures et des outils bien spécifiques et adaptés pour chaque acteur impliqué dans le programme. Ces outils permettent que les analyses contextuelles dans chaque zone soient cohérentes et contribuent à un processus de capitalisation.

En ce qui concerne l'analyse sous-jacente, VECO Benin a élaboré plusieurs documents qui permettent une analyse rigoureuse du contexte et de la problématique des stratégies choisies, concrétisées pour le Bénin dans un atelier de planification régionale. Des indicateurs sont ensuite négociés avec les acteurs concernés lors d'ateliers de travail au niveau pays. La politique et le système de suivi et d'évaluation (y inclus les outils) sont toutefois élaborés au niveau du siège en Belgique avec des adaptations au niveau du projet.

Les résultats du tableau 4 montrent qu'à une exception près, *tous les items* concernant l'analyse obtiennent des scores relativement bons tant pour la pertinence que pour l'efficacité, avec des scores légèrement plus élevés pour la pertinence que pour l'efficacité. Les différents scores peuvent être détaillés comme suit.

Les groupes cibles sont, dans l'ensemble, décrits d'une manière assez correcte et univoque (items 1.1.1 à 1.1.3). Toutefois, dans bien des cas, les groupes cibles ne sont pas clairement délimités ce qui, à l'évidence, hypothèque l'évaluabilité ultérieure. La situation (problématique) est en général bien décrite, de même que le bien fondé de l'intervention par rapport à cette situation, sans qu'il soit question pour autant de descriptions détaillées. Ceci ne semble d'ailleurs pas souhaitable compte tenu de la volonté de réduire les exigences administratives, même si cela constitue bien évidemment un handicap du point de vue de l'évaluabilité.

En ce qui concerne la description des principaux acteurs concernés en dehors du groupe cible (1.1.4), il faut remarquer que la différence entre les groupes cibles et les autres acteurs concernés n'est pas toujours claire et n'apparaît que rarement de manière explicite dans les propositions, ceci n'étant pas réellement demandé il est vrai. Du point de vue de l'évaluabilité, il serait souhaitable que l'ensemble des acteurs soit mieux décrit et que l'on ait également une description plus détaillée des groupes cibles intermédiaires et finaux, avec leur rôle. Parallèlement, il faut tenir compte du fait que certaines interventions sont très complexes, avec parfois de nombreux acteurs impliqués, ce qui rend l'analyse sous-jacente fastidieuse.

L'intégration d'une analyse genre dans l'analyse sous-jacente (1.5.5) est l'item qui, dans ce composant, obtient le score le plus faible. Ces scores faibles ne sont pas surprenants et confirment les résultats de l'évaluation réalisée récemment en ce qui concerne la sensibilité au genre de la coopération belge au développement[23]. L'absence d'une analyse genre peut fortement affaiblir l'évaluation d'une intervention, car cela empêche de faire émerger certaines raisons potentielles de l'échec d'une intervention. Plus spécifiquement, les relations entre les genres influent fortement sur le comportement des individus, si bien que ces relations ont souvent un effet sur la mise en œuvre et les résultats des interventions. Dans bien des cas, le fait d'ignorer l'influence du genre a des conséquences négatives pour la qualité des évaluations et les suggestions d'amélioration qui sont formulées. À cet égard, on constate par ailleurs que dans bon nombre d'interventions, on ne consacre que peu d'attention à la différenciation sociale au sein du groupe cible (pas seulement du point de vue du genre) et que les groupes cibles sont trop souvent représentés comme un groupe homogène. Ceci constitue un handicap pour l'évaluation de la pertinence et de l'efficacité, car la différenciation sociale peut faire en sorte qu'un projet soit plus ou moins pertinent pour certains (sous-)groupes cibles et qu'un projet, via des mécanismes sociaux qui sont trop peu identifiés, puisse aussi obtenir des scores différents sur le plan de l'efficacité par rapport à ces groupes distincts. Sur un plan plus large, ceci renforce la tendance qui consiste à ne pas envisager – ou trop peu – les effets des interventions en ce qui concerne l'égalité et l'équité (*equity*), alors qu'il s'agit tout de même de principes centraux dans la coopération au développement[24].

Les scores relativement élevés (parmi les 20% d'items avec les meilleurs scores) pour la description du lien entre l'analyse sous-jacente et les objectifs de l'intervention (1.1.6) doivent être quelque peu nuancés. On a le sentiment, surtout lorsqu'il s'agit de la deuxième ou troisième phase d'une intervention ou de programmes[25], que l'analyse sous-jacente est rédigée en fonction d'objectifs qui avaient déjà été fixés auparavant.

[23] Voir Caubergs, L., Charlier, S., Holvoet, N., Inberg, L. and Van Esbroeck, D. (2014) Un chemin difficile vers l'égalité. Évaluation du Genre et Développement dans la Coopération belge. Bruxelles : Service d'Evaluation Spécial, 135 p.

[24] Incidemment, ont peut observer que l'égalité/l'équité ne sont pas intégrés dans les cinq critères CAD classiques qui, à ce jour, sont au centre de la plupart des évaluations.

[25] Parmi les ANG, notamment, les programmes se développent bien souvent à partir des acquis du passé. Ceci n'est certainement pas négatif, mais cela comporte certains dangers comme nous le démontrons ici.

Ceci ressort entre autres du fait que l'on ne prend pas en considération différentes stratégies alternatives, mais que l'on centre l'analyse sur une seule stratégie qui est clairement établie dès le départ. En référence au cycle GCP classique, nous pourrions dire que dans de nombreux cas, il n'existe pas une véritable identification par rapport à laquelle différentes stratégies sont évaluées et l'on procède immédiatement à la formulation. Du point de vue de l'évaluabilité, ceci implique avant tout des défis supplémentaires pour l'évaluation de la pertinence qui, dans une telle situation, doit se baser sur une analyse qui n'est pas suffisamment large et/ou qui n'est pas actuelle.

Le lien entre le bien fondé de l'intervention et la politique sectorielle du pays partenaire (1.1.7) est clair dans la majorité des interventions, ce qui se traduit par les scores les plus élevés au sein de ce composant et par une place parmi les 20% d'items avec les scores globalement les plus élevés (voir aussi le tableau A3 2 de l'annexe 7)[26]. Ce score élevé indique que les acteurs responsables des interventions sont conscients de la politique sectorielle et cherchent à accorder les interventions avec cette politique, même si l'application (ou l'applicabilité) de cette politique est limitée. Vraisemblablement, l'expérience croissante au sein de la coopération belge au développement pour ce qui est de travailler avec des programmes indicatifs pluriannuels plus globaux qui visent une meilleure coordination et une plus grande concordance avec la politique (sectorielle) du pays partenaire, joue ici un rôle. Les échanges entre les acteurs belges et les plateformes existantes dans les pays partenaires peuvent aussi avoir une influence positive.

### 3.1.2 La logique d'intervention et la théorie de changement

Dans ce composant, nous avons cherché à savoir dans quelle mesure la logique d'intervention et la théorie de changement sont clairement décrites et jusqu'à quel niveau (niveau output, outcome, impact)[27]. Comme le montre ci-après l'analyse des différents items, la qualité de la logique d'intervention et de la théorie de changement est un facteur important pour l'évaluabilité. L'attention portée, dans notre cadre d'analyse, au niveau spécifique jusqu'où une logique d'intervention (ainsi que les maillons cruciaux, les hypothèses, les risques internes, etc.) est clairement et correctement développée, a des implications pour l'évaluabilité des différents critères OCDE/CAD. Autrement dit, une logique d'intervention qui est clairement et correctement développée jusqu'au niveau des outcomes mais pas jusqu'au niveau impact favorise l'évaluabilité de l'efficacité mais représente un sérieux défi pour l'évaluabilité de l'impact. Au total, 8 items ont été analysés pour ce composant (voir le tableau 5).

[26] Cette constatation, si elle ne semble pas si spectaculaire au premier abord, n'en est pas moins importante, car la majorité des items présentant les meilleurs scores appartiennent à la dimension 3 (le contexte) ; voir aussi les tableaux A3 1 et A3 2 de l'annexe 7 pour plus de détails.

[27] Ces trois niveaux correspondent aux outputs (résultats intermédiaires), à l'objectif spécifique et à l'objectif général (aux objectifs généraux) du cadre logique. Voir aussi au chapitre 2.2 le contenu qui a été donné dans cette étude aux cinq critères CAD.

**Tableau 5: Résultats majeurs en relation avec la logique d'intervention et la théorie de changement**

| | Pertinence | Efficacité | Efficience | Impact | Durabilité | Index d'évaluabilité |
|---|---|---|---|---|---|---|
| 1.2 La logique d'intervention et la théorie de changement | 2,80 | 3,00 | 3,50 | 2,10 | 3,25 | 2,96 |
| 1.2.1 Il existe une distinction claire et correcte entre les outputs et outcomes (effets) et impact | | 3,65 | | 2,65 | | 3,15 |
| 1.2.2 La théorie de changement à partir des inputs jusqu'aux outcomes et impacts finaux est clairement développée | | 3,35 | | 1,90 | | 2,63 |
| 1.2.3 La théorie de changement est logique et réaliste | | 4,15 | 4,40 | 3,20 | | 3,92 |
| 1.2.4 Des maillons critiques et cruciaux de la chaine sont identifiés et peuvent être testés | | 3,05 | 3,35 | 1,87 | | 2,76 |
| 1.2.5 La théorie de changement contient des mesures concrètes (au niveau des inputs, activités et outputs) qui doivent assurer la durabilité des bénéfices de l'intervention | | | | | 3,50 | 3,50 |
| 1.2.6 Les risques internes sont clairement signalés et analysés/ appréciés | | 3,00 | 3,10 | | 2,80 | 2,97 |
| 1.2.7 Les hypothèses externes sont clairement signalées et analysées/appréciées | 2,80 | 3,10 | 3,20 | 2,15 | 2,90 | 2,83 |
| 1.2.8 L'allocation des moyens prévus aux outputs est claire | | | 3,64 | | | 3,64 |

Le composant '*logique d'intervention et théorie de changement*' a l'un des indices d'évaluabilité les plus faibles (sixième sur huit) parmi les composants analysés (voir Tableau 3). Comme nous le montrons ci-après, le principal point névralgique est le développement de la logique d'intervention et de la théorie de changement jusqu'au niveau de l'impact. En d'autres termes, dans la plupart des interventions, l'attention se concentre avant tout sur les niveaux inférieurs et sur le développement de la théorie de mise en œuvre, tandis que les niveaux plus élevés, et plus spécialement le niveau impact (qui correspond dans notre analyse au niveau des objectifs généraux dans le cadre logique), sont souvent négligés. Or, des effets au niveau impact (notamment les effets qui découlent *directement* des outcomes) devraient déjà être visibles après la clôture d'une intervention de trois ans (soit la durée de nombreuses interventions selon le cycle de financement) et certainement après une intervention subséquente de même durée.

Les résultats du tableau 5 montrent par ailleurs que pour ce composant, le **niveau efficacité** et le **niveau efficience** sont, dans l'ensemble, bien développés. Un nombre relativement élevé d'interventions obtiennent de bons scores pour leur théorie de changement, qui est logique et réaliste jusqu'au niveau output (70%) et outcome (60%) ; cet item (1.2.3) fait d'ailleurs partie des 20% d'items avec les meilleurs scores. Par ailleurs, les maillons critiques et cruciaux sont relativement bien identifiés jusqu'à ces deux niveaux. De nombreuses interventions signalent aussi les risques internes et les hypothèses externes, même s'ils ne sont pas toujours analysés, loin s'en faut. Bien que les différences entre les pourcentages mentionnés ci-dessus pour le niveau output et le niveau outcome ne soient pas énormes, il est clair que le niveau output (efficience) est, dans l'ensemble, mieux développé que le niveau outcome (efficacité).

Le fait qu'une partie substantielle des interventions spécifie des éléments cruciaux au niveau des outcomes et surtout des outputs est positif, notamment, pour l'*apprentissage au niveau intervention*. Il est important, en particulier pour l'équipe de l'intervention directement concernée et l'organisation responsable, d'avoir un aperçu de ces éléments essentiels pour la mise en œuvre d'une intervention et qu'il est donc important de reprendre dans les interventions futures, tandis que d'autres éléments qui semblent plus accessoires peuvent être laissés de côté. L'identification des éléments critiques et cruciaux est essentielle, en particulier, dans le cas d'interventions innovantes où l'on a une connaissance insuffisante de ce qui fonctionne ou pas, et où il subsiste une incertitude quant aux modalités de mise en œuvre les plus indiquées[28]. L'identification et la vérification des hypothèses externes sont importantes, quant à elles, pour la redevabilité (*accountability*), car cela permet d'établir clairement quels facteurs exercent (ou peuvent exercer) de l'extérieur une influence sur l'intervention.

Comme nous l'avons indiqué précédemment, l'étape suivante dans la théorie de changement, à savoir le **niveau impact**, obtient des scores nettement plus faibles. Plus précisément, les éléments les plus faibles au sein de la logique d'intervention et de la théorie de changement concernent en particulier le critère de l'impact. Dans un nombre relativement élevé d'interventions, la théorie de changement n'est pas développée jusqu'au niveau impact et lorsque le niveau impact est spécifié, il se situe souvent à une grande distance des outcomes, si bien qu'il se crée un *'missing middle'* et qu'il devient impossible de déterminer la contribution d'une intervention individuelle jusqu'à cet impact[29]. Cet effet est renforcé par le fait que, bien souvent, il n'y a pas non plus de maillons critiques et cruciaux ni d'hypothèses externes qui sont définis entre le niveau outcome et le niveau impact (le score de cet item (1.2.4) fait partie des 20% de scores les plus faibles). Il est clair que ceci compromet la possibilité d'obtenir des évaluations d'impact (et la qualité de telles évaluations), étant donné que c'est précisément l'analyse de ces éléments qui permet une différenciation entre les problèmes au niveau de la mise en œuvre, les lacunes dans la théorie de changement sous-jacente et les influences d'autres facteurs externes sur lesquels l'intervention n'a pas de contrôle. En outre, c'est surtout en identifiant et en testant les maillons critiques et cruciaux que l'on peut arriver à cerner les raisons pour lesquelles une intervention génère ou non un impact. Ceci affaiblit pour une bonne part la fonction d'apprentissage/rétroaction au niveau supérieur de l'intervention qui doit fournir, notamment aux décideurs (et à d'autres acteurs qui ne sont pas directement impliqués dans la mise en œuvre d'interventions spécifiques), des informations intéressantes en ce qui concerne l'outcome et l'impact. Enfin, l'élaboration d'une TdC jusqu'au niveau impact peut aussi fournir des informations sur les modalités de mise en œuvre qui sont les plus efficaces et efficientes et qui offrent le plus de chances quant à la réalisation de l'impact (sans pour autant qu'une évaluation de l'impact doive effectivement être réalisée à chaque fois).

Les scores faibles pour le critère 'impact' peuvent sans doute s'expliquer en partie par le fait qu'il y a – en particulier pour les petites interventions ou organisations, mais aussi en général – peu d'*incitants* à dépasser le niveau de la mise en œuvre : le développement d'une logique d'intervention au niveau de la mise en œuvre a bien plus d'utilité directe pour l'intervention/organisation même et c'est le niveau sur lequel l'intervention/organisation a le plus de contrôle. D'autre part, les liens avec les niveaux supérieurs sont souvent flous et moins contrôlables, l'identification des éléments critiques et des maillons cruciaux aux niveaux supérieurs est plus difficile, et cela nécessite souvent une analyse sectorielle et contextuelle plus poussée. En outre, jusqu'il y a peu, les bailleurs de fonds attachaient relativement peu d'importance à la théorie de l'impact. Ceci ressort entre autres du fait que les canevas des propositions de projets

[28] Ceci se rapproche quelque peu de l'idée de *'structured experiential learning'*, voir Pritchett, L., Samji, S. and Hammer, H. (2013) "It's all about MeE: using structured experiential learning to crawl the design space", *Center for Global Development Working Paper 322*. Washington, D.C.: Centre for Gobal Development.

[29] Il existe des solutions pour y remédier, par exemple insérer un niveau supplémentaire dans la théorie de changement et travailler avec des 'intermediate outcomes' et des 'final outcomes' afin de mettre en place un échelon intermédiaire entre outcomes et impact. Nous y reviendrons dans les recommandations.

permettent que l'on n'accorde que peu ou pas d'attention à ce niveau ; de même, dans la grille d'appréciation de la DGD pour les actions dans le Sud, l'impact ne fait l'objet d'aucune attention, sans doute parce que les bailleurs de fonds restent avant tout intéressés par la '*redevabilité*' et sont conscients par ailleurs que les interventions n'ont (ne peuvent avoir) le contrôle que sur les activités et les outputs. Comme nous le montrons plus loin dans ce rapport, l'absence d'une théorie de changement entièrement élaborée a aussi une influence sur le système S&E (tant sur papier que dans la pratique), et cela met notamment l'évaluabilité de l'impact sous forte pression.

Si la manière dont les interventions sont élaborées a un effet négatif sur l'évaluabilité de l'impact, les scores pour la **durabilité** sont un peu plus positifs. La plupart des interventions formulent des mesures concrètes – et même des outputs concrets – qui doivent garantir la pérennité des bénéfices de l'intervention. Ceci reflète le souci croissant des organisations de préserver les résultats de leurs interventions une fois celles-ci achevées. Lors de l'appréciation des risques internes et des hypothèses externes, il est cependant beaucoup moins tenu compte de la durabilité : 28% des interventions ont signalé et analysé des risques internes liés à la durabilité, tandis que 18% seulement l'ont fait pour des hypothèses externes.

Enfin, il apparaît que toutes les interventions fournissent des informations en ce qui concerne les moyens prévus et que dans 10% seulement des interventions, l'allocation des moyens/frais prévus aux outputs n'est absolument pas claire. Le bon score pour cet item a un effet positif sur – et est en fait crucial pour – l'évaluabilité de l'efficacité et est en partie attribuable à l'attention croissante accordée à ce critère dans les formats financiers imposés par la DGD.

### 3.1.3 Le système S&E proposé

Sous ce composant, nous avons cherché à savoir dans quelle mesure le plan d'intervention fournit des informations en ce qui concerne le système S&E. La qualité du système S&E (proposé) constitue, pour des raisons évidentes, un important facteur d'évaluabilité : plus ce système est développé, plus grande est sa capacité à générer des informations qui seront profitables à l'évaluabilité.

L'analyse des neuf items sous ce composant nous a permis d'obtenir un aperçu du système S&E proposé. Dans ce cadre, il est important de garder à l'esprit que l'aperçu du système S&E *proposé* ne nous apprend rien quant à (la qualité de) l'application *réelle* de ce système (cette application réelle est abordée à la section 2.2).

Les principaux résultats sont présentés succinctement sous forme de tableau ci-dessous (tableau 6), de la même manière que pour les deux composants précédents.

**Tableau 6: Résultats majeurs quant au système S&E proposé**

| | Pertinence | Efficacité | Efficience | Impact | Durabilité | Index d'évaluabilité |
|---|---|---|---|---|---|---|
| 1.3 Le système S&E proposé | 2,45 | 2,83 | 2,85 | 1,95 | 2,46 | 2,51 |
| 1.3.1 Les résultats majeurs envisagés de l'intervention sont bien opérationnalisés | | 3,15 | 3,65 | 1,83 | 2,72 | 2,84 |
| 1.3.2 Là où cela est nécessaire/ pertinent, les indicateurs sont spécifiés selon le sexe ou selon un autre paramètre pertinent | | 1,82 | 1,81 | 1,34 | 1,63 | 1,65 |
| 1.3.3 Le système S&E proposé contient une opérationnalisation consistante de la logique d'intervention et la théorie de changement sous-jacentes | 3,00 | 3,45 | 3,65 | 2,30 | 2,69 | 3,02 |
| 1.3.4 L'approche de suivi et d'évaluation de la réalisation des résultats de l'intervention et de leur durabilité est clairement décrite | 2,70 | 3,15 | 3,28 | 2,13 | 2,62 | 2,78 |
| 1.3.5 L'approche de suivi des hypothèses est clairement décrite | 1,85 | 1,85 | 1,90 | 1,55 | 1,82 | 1,79 |
| 1.3.6 L'approche de suivi des risques internes est clairement décrite | 1,70 | 1,80 | 1,85 | 1,45 | 1,68 | 1,70 |
| 1.3.7 Les ressources personnelles et financières du système S&E sont clairement décrites | | | 2,40 | | | 2,40 |
| 1.3.8 Le système MIS (système d'information de gestion) permet une allocation des dépenses aux outputs et composantes spécifiques de l'intervention | | | 3,67 | | | 3,67 |
| 1.3.9 La façon selon laquelle le système S&E s'articule par rapport aux systèmes de S&E national/local est bien décrite | 1,57 | 1,69 | 1,63 | 1,57 | 1,57 | 1,61 |

Le système S&E proposé est le composant qui présente l'indice d'évaluabilité le plus faible parmi les huit composants analysés, et ce pour les cinq critères d'évaluation CAD ; sept (!) des neuf items de ce composant font partie – au moins pour un critère et souvent pour tous les critères – du groupe des 20% d'items avec les scores les plus faibles (voir aussi le tableau A3 3 à l'annexe 7). Du point de vue de l'évaluabilité théorique, ceci ne constitue pas un véritable problème *en soi*. Ex ante, ce sont avant tout la qualité de l'analyse sous-jacente et celle de la logique d'intervention et de la théorie de changement qui sont importantes pour l'évaluabilité. La qualité du système S&E est surtout importante du point de vue de l'évaluabilité pratique, mais dans ce cas il s'agit plutôt de la *pratique* S&E et pas tellement du système S&E sur papier, tel qu'il est présenté dans la proposition d'intervention. Il existe toutefois pour quatre critères, comme nous le montrons plus loin, une forte corrélation entre la qualité du plan S&E tel qu'il est présenté dans la proposition d'intervention et la pratique S&E au final (voir l'analyse du composant 2.2 ci-après). Dans cette optique, (la qualité de) l'élaboration d'un système S&E dans la phase initiale est un bon indicateur de la pratique ultérieure.

Il peut y avoir plusieurs explications à la faiblesse du score de ce composant. Premièrement, les exigences du bailleur de fonds en ce qui concerne la description du système S&E dans la proposition d'intervention ne sont pas très élaborées ; à l'évidence, une plus grande attention est demandée – à tort ou à raison – pour d'autres aspects. Il en résulte, vraisemblablement, que les promoteurs des interventions investissent

relativement peu dans ce composant. Cette hypothèse est confirmée par le fait que la pratique S&E proprement dite obtient des scores sensiblement (environ 20%) plus élevés (voir tableau 3). D'autre part, bon nombre de promoteurs n'ont pas l'habitude de décrire en détail leur système S&E. Un autre élément qui intervient est que, d'un point de vue mental, on ne commence véritablement à accorder de l'attention au S&E qu'une fois la mise en œuvre lancée ; comme nous le verrons plus loin, il y a des interventions qui ne commencent réellement à définir et à développer leur système S&E qu'au moment du démarrage. Un dernier élément est qu'au moment de l'étude (printemps 2015), certaines organisations n'avaient développé que récemment une stratégie S&E (ou s'y affairaient encore), si bien que l'opérationnalisation sur le terrain a pris du retard.

Lorsque nous examinons les cinq critères d'évaluation du CAD, il apparaît – sans surprise – que 'l'impact' enregistre le score le plus faible, tandis que l'efficacité et l'efficience ont le meilleur score, même s'il faut préciser que le score, pour ces deux critères, reste encore inférieur au score moyen de 3. L'explication du score relativement plus élevé pour l'efficacité et l'efficience réside bien entendu dans le fait que les systèmes S&E sont centrés avant tout sur l'objectif spécifique à atteindre, et en particulier sur les outputs à réaliser, et que – par corollaire – dans la politique S&E de la plupart des organisations, une plus grande importance est accordée au suivi par rapport à l'évaluation. Par ailleurs, le peu d'attention accordée à l'impact dans la logique d'intervention et la théorie de changement (voir l'analyse du composant 1.2 ci-dessus) joue aussi un rôle important. Enfin, si les scores ne sont pas vraiment bons, c'est également lié au fait que dans un certain nombre de cas, les compétences de base en ce qui concerne la formulation de bons indicateurs ne sont, à l'évidence, pas suffisamment maîtrisées[30].

L'affirmation selon laquelle les systèmes S&E sont axés avant tout sur les outputs et les activités et moyens sous-jacents est aussi confirmée par le score relativement bon, au niveau de l'efficience, en ce qui concerne l'opérationnalisation des résultats visés (1.3.1). D'un autre côté, le score pour l'impact est ici très faible, principalement en raison du fait que bon nombre de propositions n'accordent pas la moindre attention à l'impact. Le système S&E proposé forme, en ce sens, une préfiguration de la pratique S&E (voir le chapitre 2.2 ci-après) où, comme nous le constaterons, les systèmes S&E sont relativement bien développés au niveau opérationnel (suivi au sens strict), mais leur qualité et leur portée diminuent à mesure qu'on monte dans la chaîne moyens-fin et qu'on doit élargir le champ de vision.

De même, en ce qui concerne la traduction – dans le système S&E – de la logique d'intervention et de la théorie de changement sous-jacentes (item 1.3.3), c'est sur le plan de l'efficience et de l'efficacité que les scores sont les meilleurs. Il apparaît également que les systèmes S&E, jusqu'à un certain niveau, accordent une certaine attention aux activités et outputs qui doivent favoriser la durabilité des interventions (item 1.3.4) ; les scores en question ne sont pas vraiment bons mais témoignent malgré tout d'une attention explicite pour la durabilité dans de nombreuses interventions. Ici aussi, nous pouvons déceler un parallèle évident avec les résultats relatifs à la durabilité dans l'analyse du composant précédent (voir l'analyse du composant 1.2 ci-dessus).

Les scores (très) faibles en ce qui concerne l'utilisation d'indicateurs spécifiques au sexe (et d'autres facteurs de différenciation sociale) (1.3.2) et le suivi des hypothèses externes (1.3.5) s'expliquent, jusqu'à un certain point, par le manque d'attention pour ces points dans l'analyse sous-jacente et le développement de la logique d'intervention et de la théorie de changement (voir l'analyse des composants 1.1 et 1.2 ci-dessus). Manifestement, l'hypothèse selon laquelle, si ces éléments ne sont pas pris en compte dès le début (par exemple via les données et risques 'baseline' spécifiques au sexe), ils interviennent (trop) peu par la suite, se vérifie ici aussi. D'un autre côté, l'inverse s'applique également, comme l'illustre la *bonne pratique* dans le Cadre 2.

[30] Il est notamment fort peu tenu compte de l'exigence qui veut que les indicateurs doivent être 'spécifiques' (le 'S' de SMART) : l'indicateur doit en effet renvoyer de manière 'spécifique' au résultat visé par le biais de l'objectif (output, outcome, impact).

*Cadre 2: Analyse des risques dans le projet PASAB II (Caritas)*

Dans les documents de l'intervention PASAB II de l'ONG Caritas Rwanda, on identifie par résultat des risques qui peuvent potentiellement influencer les résultats des activités prévues. On fait aussi une analyse des risques où on identifie pour chaque risque les effets potentiels et les actions qu'on va mener si les risques se réalisent et on assigne les responsabilités et la périodicité de contrôle. Les résultats sont suivis mais aussi les risques : il y a un document dans lequel on compare les risques prévus avec la situation actuelle. Ce document contient de l'information intéressante qui peut aussi être utilisée au moment d'une évaluation car ces informations donnent une idée des facteurs qui ont influencé les résultats. Les compétences en suivi évaluation des activités sont transférées aux structures et agents œuvrant dans la communauté (paysans relais et paysans de contact) qui, de leur côté, assurent le suivi indépendamment de l'appui du projet, ce qui assure la pérennisation des acquis du projet. La combinaison zone pilote, paysans relais, paysans de contact, structure paysanne et équipe du projet constitue un support important dans la transmission de l'information fiable.

Le score également très faible en ce qui concerne l'approche de suivi des risques internes (1.3.6) est lié, selon nous, au fait que peu d'organisations sont enclines à décrire ces risques dans leurs propositions d'intervention (voir aussi l'analyse du composant 1.2 ci-dessus). Ce point de vue est compréhensible en ce sens que les organisations savent que leurs propositions seront passées à la loupe et veulent dès lors éviter de se montrer faibles. Il y a par ailleurs des organisations qui soutiennent que ces risques font partie de leur cuisine interne, n'ont pas de lien direct avec la qualité et la mise en œuvre des programmes et ne sont dès lors pas – ou ne doivent pas être – communiqués. Ces arguments semblent fondés si l'on compare les scores avec la qualité de l'information relative à la pratique de l'analyse des risques, qui donne un meilleur score (voir plus loin, item 2.1.9). D'un autre côté, du point de vue de l'évaluabilité, il est souhaitable et important que des informations soient disponibles sur ces risques internes, la manière dont on entend gérer ces risques et la manière dont ils seront suivis.

Les ressources personnelles et financières qui sont disponibles pour le système S&E ne sont, dans l'ensemble, pas décrites ou le sont de façon très limitée (1.3.7). Nous ne trouvons des informations substantielles sur ce point que dans un quart des propositions. Dans bien des cas, par exemple, les évaluations planifiées, même si elles sont d'une certaine ampleur, ne sont pas mentionnées spécifiquement dans le budget. Ici aussi, il semble que l'on ne juge pas ces informations suffisamment cruciales pour être reprises dans une proposition d'intervention. Du point de vue de l'évaluabilité, cela implique que l'on a une vision limitée de l'importance relative qui est accordée *ex ante* au S&E.

La façon selon laquelle le système S&E de l'intervention s'articule par rapport au système S&E national/local n'est, dans la plupart des cas, pas clairement décrite comme le montre le score très faible pour cet item (1.3.9). Ce faible score s'explique en partie par le fait qu'une série d'interventions sont de taille très réduite ou s'occupent de choses qui, au niveau (administratif) supérieur, ne sont pas intégrées dans des systèmes S&E. D'autres facteurs jouent également, comme l'absence, la fiabilité douteuse ou le non-fonctionnement de tels systèmes sur le plan national et décentralisé (par exemple en RDC), si bien qu'ils ne sont pas jugés adéquats pour satisfaire à la redevabilité (*accountability*) ascendante vis-à-vis de l'autorité qui assure le financement. Toutefois, il est clair également que trop peu d'interventions vérifient effectivement dans quelle mesure leur système S&E, et en particulier le suivi des indicateurs clés au niveau outcome et impact, pourrait s'accorder avec ce qui se fait au niveau supérieur et au niveau décentralisé (et éventuellement au sein des organisations partenaires)[31].

Le score le plus élevé dans ce composant est celui de l'item "*Le système MIS permet une allocation des dépenses aux outputs et composantes spécifiques de l'intervention*" (1.3.8). Du point de vue de l'évaluabilité, ceci est un acquis important qui est lié à la manière dont le bailleur de fonds souhaite que les budgets soient établis et à l'importance qui est accordée à un budget bien élaboré dans la proposition. En imposant

[31] Nous reviendrons sur les conséquences de ce manque au point 2.2. de cette partie.

un modèle qui relie les différentes catégories de dépenses aux outputs, on donne accès à des informations importantes pour une analyse de l'efficience.

*Last but not least*, il est important d'observer que dans peu d'interventions, il est question d'un véritable *système* S&E. On trouve bien, dans la plupart des propositions, des informations relatives aux composantes d'un tel système, mais elles ne sont pas – ou en partie seulement – accordées entre elles et regroupées en un ensemble cohérent. Ainsi, il arrive régulièrement que l'analyse des risques (risques internes et hypothèses externes) mentionne des facteurs importants que l'on ne retrouve pas dans le cadre logique, et inversement. Par ailleurs, l'absence d'un véritable *système* se répercute dans la pratique S&E.

### 3.1.4 La consistance et l'adaptation de la logique d'intervention et de la théorie de changement

Pour l'évaluabilité des interventions, il est essentiel que les éventuels changements dans la logique d'intervention et dans la théorie de changement sous-jacente soient *clairement* indiqués et intégrés dans le système S&E. Il est important également que l'on spécifique clairement pourquoi et comment ces changements ont été mis en œuvre. Si les changements ne sont pas (clairement) indiqués et intégrés dans le système S&E, certaines informations pertinentes concernant la situation réelle feront défaut, ce qui peut conduire à des conclusions incorrectes sur la qualité des interventions (en lien avec tous les critères OCDE/CAD) et les facteurs qui influent sur la mise en œuvre et les effets des interventions. Si le système S&E n'est pas adapté, cela implique aussi qu'une partie de l'intervention rénovée n'est pas suivie.

Dans un quart environ des interventions, des changements sont opérés dans la logique d'intervention et la théorie de changement sous-jacente lors de la mise en œuvre[32]. La majorité de ces adaptations se limitent au niveau output et aux niveaux inférieurs.

**Tableau 7: Résultats majeurs quant à la consistance et adaptation de la logique d'intervention et la théorie de changement**

| | Pertinence | Efficacité | Efficience | Impact | Durabilité | Index d'évaluabilité |
|---|---|---|---|---|---|---|
| 1.4 La consistance et l'adaptation de la logique d'intervention et la théorie de changement | 3,55 | 3,44 | 3,52 | 3,50 | 3,75 | 3,54 |
| 1.4.1 Des changements éventuels dans la logique d'intervention et la théorie de changement sous-jacente sont bien indiqués et argumentés | 4,27 | 4,00 | 3,76 | 4,20 | 4,17 | 4,03 |
| 1.4.2 L'information est disponible sur la vision et les opinions des parties concernées les plus importantes quant aux changements éventuels dans la logique d'intervention et la théorie de changement | 2,82 | 2,63 | 2,43 | 2,80 | 2,67 | 2,63 |
| 1.4.3 Des changements éventuels dans la logique d'intervention et la théorie de changement sont intégrés de façon adéquate dans le système S&E | 3,18 | 3,25 | 3,10 | 3,00 | 3,33 | 3,17 |

[32] Le nombre limité d'interventions qui ont pu être analysées ici incite à la prudence quant aux résultats.

Ce composant obtient un score relativement bon comparativement aux autres composants de la dimension 1 et de la dimension 2 (voir le tableau 7). Les scores pour les trois items présentent toutefois une grande variation, avec des scores relativement élevés pour l'indication et l'argumentation des changements éventuels dans la logique d'intervention et la théorie de changement sous-jacente (dont certains figurent parmi les 20% de scores les plus élevés) et des scores relativement faibles pour la disponibilité de l'information sur la vision et les opinions des parties concernées les plus importantes quant aux changements éventuels.

Le fait que les changements sont indiqués dans la plupart des interventions est en soi une constatation positive qui illustre la qualité de la gestion et, parallèlement, la volonté de documenter et d'argumenter les changements. Si les scores pour l'intégration adéquate des changements dans le système S&E sont corrects, le fait qu'un nombre important d'interventions où des changements ont été opérés (et indiqués) ne les ont pas répercutés dans le système S&E est néanmoins préoccupant. Les scores individuels faibles pour cet item vont souvent de pair avec des scores faibles pour l'ensemble du système S&E : là où les systèmes sont faibles, les changements ne sont généralement pas appliqués dans le système S&E.

## 3.2 Analyse de la pratique de mise en œuvre et de gestion de l'intervention et du contexte

Dans cette partie sont passés en revue deux composants très élaborés : l'information de base (et la disponibilité de cette information) en ce qui concerne la mise en œuvre de l'intervention (11 items) et le système S&E dans la pratique (12 items).

### 3.2.1 La disponibilité de l'information de base quant à la mise en œuvre de l'intervention

Sous ce composant, nous avons examiné si l'information de base relative à la mise en œuvre de l'intervention était présente. Ceci concerne aussi bien l'information qui devrait être présente au début d'une intervention (p. ex. proposition d'intervention, baseline) que l'information relative à la progression de l'intervention. La présence de l'information de base quant à la mise en œuvre de l'information est essentielle pour l'évaluabilité, étant donné que sans cette information, il est difficile d'établir une comparaison entre la situation initiale, la situation intermédiaire et la situation finale, si bien que les progrès dans la mise en œuvre et les effets d'une intervention sont difficiles à établir, et donc à évaluer.

**Tableau 8: Résultats majeurs en relation avec la disponibilité de l'information de base quant à la mise en oeuvre de l'intervention**

| | Pertinence | Efficacité | Efficience | Impact | Durabilité | Index d'évaluabilité |
|---|---|---|---|---|---|---|
| 2.1 Disponibilité de l'information de base quant à la mise en oeuvre de l'intervention | 2,95 | 3,20 | 3,53 | 2,13 | 2,62 | 2,88 |
| 2.1.1 Les documents de base (*attendus*) sont disponibles | 4,03 | 4,05 | 4,13 | 3,73 | 3,95 | 3,97 |
| 2.1.2 L'information de base (en ligne avec la logique d'intervention) concernant le groupe cible est disponible | | 3,50 | 3,90 | 2,45 | 2,84 | 3,18 |
| 2.1.3 L'information de base (en ligne avec la logique d'intervention) concernant le counterfactual est disponible | | 1,28 | 1,28 | 1,44 | 1,29 | 1,32 |
| 2.1.4 L'information de base est disponible quant aux indicateurs pertinents est spécifiée selon le genre et/ou autres paramètres pertinents | | 1,76 | 1,86 | 1,51 | | 1,71 |
| 2.1.5 L'information (attendue) est disponible quant au progrès vers la réalisation des objectifs de l'intervention | | 3,37 | 3,93 | 1,69 | | 3,04 |
| 2.1.6 L'information quant à la participation du groupe cible initial est disponible | 3,26 | 3,31 | 3,41 | 2,84 | 3,11 | 3,19 |
| 2.1.7 L'information quant à la collecte de données liées aux indicateurs est disponible | 2,73 | 3,08 | 3,28 | 2,28 | 2,69 | 2,82 |
| 2.1.8 Le plan/la proposition de collecte de données permet en principe une collecte fiable des données en relation avec les indicateurs | 3,00 | 3,12 | 3,47 | 2,52 | 3,00 | 3,04 |
| 2.1.9 L'information est disponible quant au suivi des | | 2,40 | 2,75 | 2,13 | 2,86 | 2,54 |

| | Pertinence | Efficacité | Efficience | Impact | Durabilité | Index d'évaluabilité |
|---|---|---|---|---|---|---|
| risques internes et les conséquences éventuelles pour la logique d'intervention et la mise en oeuvre du projet sont indiquées | | | | | | |
| 2.1.10 L'information est disponible quant au suivi des hypothèses et les conséquences éventuelles pour la logique d'intervention et la mise en oeuvre du projet sont indiquées | 2,21 | 2,37 | 2,60 | 2,20 | 2,45 | 2,37 |
| 2.1.11 Les dépenses de l'intervention sont bien enregistrées/ documentées et peuvent connectées aux outputs | | | 3,95 | | | 3,95 |

Le composant '*Information de base disponible quant à la mise en œuvre de l'intervention*' présente, à une exception près, le score le plus bas des 8 composants examinés quant à l'indice d'évaluabilité (voir tableau 3). Ceci est quelque peu surprenant dans la mesure où l'on peut supposer que du point de vue du S&E, "avoir une information disponible" est la toute première étape. Comme le montre l'analyse des différents items, il y a une grande variation dans les scores, avec avant tout des scores très bas pour la présence d'une information de base concernant un 'counterfactual' (ce qui complique surtout l'évaluabilité de l'impact) et pour la disponibilité d'une information décomposée selon le genre ou autres paramètres pertinents. Une constatation plus nuancée est qu'une grande quantité d'information est bel et bien collectée et disponible, mais que cette information concerne surtout les niveaux inférieurs de la chaîne et n'a que peu de rapport, d'une part avec les éléments qui sont *extérieurs à l'intervention* (comme le counterfactual, les hypothèses externes, etc.), d'autre part avec le processus de collecte de données proprement dit (type de méta-informations). Cette constatation est en grande partie liée à des éléments observés précédemment au sujet des composants relatifs à la logique d'intervention (voir 1.2) et au système S&E proposé (voir 1.3), où il était clair déjà que l'attention se porte essentiellement sur le niveau opérationnel et le niveau de la mise en œuvre. Le fait qu'il n'existe pas de tradition ni d'exigences clairement formulées quant aux informations à fournir sur l'organisation concrète du S&E (mode de collecte et de traitement des données) peut jouer également.

Comme le montre le tableau 8, l'item avec les meilleurs scores dans ce composant est "la disponibilité de documents de base" (2.1.1)[33]. Les documents de base comprennent entre autres les propositions d'intervention, le dossier technique et financier, les rapports 'baseline' et les rapports de mise en oeuvre. Jusqu'au niveau des outcomes, ces documents contiennent assez souvent des informations sur l'avancement de la réalisation des objectifs de l'intervention (2.1.5), une comparaison étant établie avec la situation de départ. Cette dernière information est importante pour pouvoir franchir pas de plus au-delà du suivi. Pour des raisons évidentes, cette information est par contre fort peu présente pour le niveau impact. Le bon score en ce qui concerne la présence des documents de base doit cependant être quelque peu nuancé, car dans certains cas, l'information n'est présente que dans une langue qu'une partie des acteurs ne maîtrise pas. D'autre part, les documents disponibles ne sont pas toujours connus de tous les acteurs, ce qui implique une appropriation limitée du contenu de ces documents parmi ces acteurs. Ceci a une influence sur l'évaluabilité d'une intervention, car une évaluation se base, dans ces cas-là, sur des informations qui ne sont pas connues de tous les

[33] … même si, dans certains cas, les parties concernées ont dû être sommées à plusieurs reprises de fournir ces informations. En outre, il est important de signaler que la qualité de ces documents varie fortement, ce dont témoignent les scores des autres items.

acteurs et la disponibilité ou non de certaines informations dépend donc des acteurs spécifiques qui sont impliqués dans l'évaluation[34].

Deux items affichent des scores exceptionnellement faibles pour les cinq critères d'évaluation CAD : la disponibilité de l'information de base concernant le counterfactual (2.1.3) et la spécification selon le genre ou autres paramètres des indicateurs pertinents dans l'information de base (2.1.4) ; ces deux items figurent en bonne place dans la liste des 20% d'items avec les scores les plus faibles (voir aussi le tableau A3 3 de l'annexe 7). L'absence d'informations désagrégées n'est pas une surprise et doit être mise en relation avec l'absence d'une analyse de genre et d'indicateurs désagrégés dans le système S&E proposé (voir plus haut).

Le 'counterfactual' est une estimation de la situation qui se présenterait si l'intervention n'avait pas eu lieu et est surtout important pour la mise en œuvre correcte, sur le plan méthodologique, des évaluations d'impact[35]. Les mauvais scores pour cet item sont dus en partie au fait qu'il n'est pas toujours simple d'identifier un bon counterfactual, que ceci nécessite certaines connaissances méthodologiques et qu'il s'agit d'une pratique qui n'est pas encore entrée dans les mœurs dans la coopération belge au développement. Par ailleurs, l'utilisation d'un counterfactual semble être souvent interprétée d'une manière très limitative et se réduire à l'utilisation d'un groupe de contrôle de type RCT (*Random Controlled Trial*). Cette piste, il est vrai, n'est pas toujours possible ou même souhaitable, mais il existe de nombreuses alternatives qui sont souvent moins complexes mais aussi, à l'évidence, moins connues[36]. Citons, à titre d'exemple, la sélection d'un groupe non-intervention par 'matching', la sélection d'un groupe comparable impliqué dans une autre intervention (pour mieux cerner l'impact différentiel), l'utilisation de contrôles statistiques, de contrôles génériques et de types plus complexes de comparaisons 'avant-après' tels que panels, time series ou shadow controls[37]. Quelques

---

[34] Pour cette étude, cela signifie également que les scores attribués sur la base de l'étude documentaire préparatoire en ce qui concerne cet item ont dû, pour de très nombreuses interventions, être adaptés au cours de la mission sur le terrain. Si un évaluateur ne peut pas effectuer de visite de terrain préparatoire, cela a donc d'importantes implications quant à savoir s'il peut suffisamment se préparer à la mission sur le terrain et indirectement à l'évaluabilité.

[35] 'Impact' est utilisé ici dans le sens des effets directement attribuables à l'intervention (causalité) (voir aussi la note de bas de page 14), qui peuvent donc se situer aussi bien au niveau des 'outcomes' qu'au niveau des maillons situés plus haut dans la chaîne.

[36] Voir Rossi P.H., Lipsey M.W. and H.E. Freeman (2004) *Evaluation: a systematic approach*, 7th edition. Thousand Oaks: Sage; Bamberger M., J. Rugh, M. Church and L. Fort (2004) "Shoestring evaluation: designing impact evaluations under budget, time and data constraints", *American Journal of Evaluation* 25 (1): 5-37.

[37] Dans le cas du *'matching'*, on recherche un groupe de contrôle qui soit similaire au groupe d'intervention au niveau des caractéristiques dont on présume qu'elles ont une influence sur les résultats de l'intervention. Dans les *contrôles statistiques*, on suit une logique similaire, mais dans ce cas, le contrôle ne porte pas sur certaines caractéristiques au moment de l'organisation de l'évaluation et de la collecte des données, mais bien au moment de l'analyse des données (autrement dit, lors de l'organisation de l'évaluation, on ne recherche pas un groupe similaire mais on contrôle par rapport aux caractéristiques déterminantes en les intégrant dans une régression en tant que variables de contrôle, ce qui permet d'écarter l'influence de ces variables). L'inconvénient, dans les counterfactuals précités, est qu'ils nécessitent une collecte de données dans un groupe qui n'est pas impliqué dans l'intervention, ce qui n'est pas toujours possible ou souhaitable. Dans le cas des *contrôles génériques*, on utilise des données disponibles au niveau national quant au résultat que l'on souhaite atteindre avec une intervention spécifique. Autrement dit, on utilise une population entière (ou la population d'une région) comme groupe de contrôle pour le groupe d'intervention. Les contrôles génériques sont évidemment l'option la moins coûteuse, mais ils ne sont pas toujours disponibles (ils le sont en général pour les secteurs enseignement, soins de santé, etc.) et bien souvent, ne sont pas assez affinés du fait que les scores moyens pour une population étendue ne peuvent, dans bien des cas, être considérés comme étant représentatifs pour le groupe d'intervention spécifique (que l'on choisit précisément pour certaines caractéristiques distinctives par rapport à une population moyenne). Les formes plus sophistiquées de comparaison 'avant/après' sont une autre possibilité. En principe, une comparaison 'avant/après' ne permet pas de donner une bonne indication de l'impact (lequel requiert une comparaison 'avec/sans'). Toutefois, en effectuant plus de mesures de la situation avant et après, on peut renforcer la comparaison avant/après : dans le cas du *panel*, on suit une série de ménages, tout au long de l'intervention et après ; dans le cas des *times series*, on utilise au minimum 30 observations pour l'intervention et on extrapole sur cette base pour se faire une idée

interventions dans l'échantillon ont utilisé un counterfactual qui n'était ni complexe ni coûteux, et plusieurs autres pourraient identifier de manière relativement simple un counterfactual.

Dans la plupart des interventions, il y a une information de base concernant le (la participation du) groupe cible (2.1.2 et 2.1.6), mais en général, cette information est incomplète et se limite à ceux qui sont effectivement atteints. Dans la proposition d'intervention déjà, l'information concernant le (la délimitation du) groupe cible est souvent très générale, si bien que cette information reste souvent très quantitative et fournit peu d'informations qualitatives sur la situation dans laquelle se trouve le groupe cible. C'est quelque peu surprenant, car souvent des études de base ont bien été réalisées, mais elles esquissent une image assez générale de la région ou du secteur et ne donnent pas une analyse approfondie des caractéristiques spécifiques du groupe cible (et d'un counterfactual), ni des facteurs susceptibles d'influencer la mise en œuvre et les résultats de l'intervention. C'est précisément une telle analyse de base qui permet d'identifier les paramètres (comme le genre, entre autres) qui peuvent avoir une influence sur la mise en œuvre et les résultats d'une intervention et qu'il convient par conséquent de reprendre dans un système S&E, étant donné qu'ils fournissent des informations cruciales qui influent fortement sur la qualité et l'exhaustivité de l'évaluation.

La présence d'une information de base désagrégée aussi large que possible est essentielle pour l'évaluabilité des interventions, étant donné que cela permet une comparaison entre la situation initiale et la situation finale. Une information spécifique concernant le groupe cible effectivement atteint en comparaison avec le groupe cible défini initialement est, quant à elle, essentielle pour déterminer si l'intervention atteint son groupe cible initial. Cette information est une première étape dans l'analyse de la manière avec laquelle l'intervention atteint ou non son groupe cible et permet d'évaluer si l'intervention est conçue de telle sorte que le groupe cible prédéfini puisse effectivement être atteint. Par ailleurs, le manque d'informations en ce qui concerne les 'drop-outs' peut facilement conduire à une surestimation des effets des interventions.

Les scores pour les items relatifs au suivi des risques internes et des hypothèses (2.1.9 et 2.1.10) sont médiocres à faibles pour tous les critères (ils font tous deux partie des 20% d'items avec les scores les plus faibles), avec dans l'ensemble un score un peu meilleur pour le suivi des risques internes que pour le suivi des hypothèses externes. D'autre part, le niveau output enregistre un score un peu plus élevé que le niveau outcome, et ce dernier plus élevé que le niveau impact. Cette constatation n'est pas surprenante et est liée à l'attention – comme indiqué précédemment – accordée avant tout au niveau mise en œuvre et aux éléments qui sont directement en lien avec l'intervention.

Pour la plupart des interventions, des informations sont présentes jusqu'au niveau outcome en ce qui concerne la collecte de données liées aux indicateurs (2.1.7), mais seules quelques interventions présentent des informations complètes et contenant, entre autres, des détails sur ce qui est collecté, par qui, avec quelle fréquence et quelle couverture et avec quelles méthodes de collecte de données (voir le cadre 3 pour un exemple de bonne pratique). Par ailleurs, l'information en ce qui concerne la collecte de données est souvent très fragmentaire, dispersée à la fois entre plusieurs documents et entre différents chapitres d'un même document. Une telle information sur le processus de collecte de données est importante pour l'évaluabilité des interventions, car cette

---

de la situation après l'intervention (= sans intervention). On compare cette information avec la situation réelle après l'intervention (= avec l'intervention), ce qui donne une bonne idée de l'impact réel. On peut opter, enfin, pour des "shadow controls" : dans ce cas, l'évaluateur tentera, sur la base de sa propre expérience et d'entretiens avec des experts locaux du secteur, avec des bénéficiaires, etc., de faire une évaluation aussi précise que possible de ce que serait la situation sans l'intervention. C'est la méthode la moins scientifique, mais plus les sources d'information et les connaissances sur lesquelles on s'appuie sont solides, plus ce 'counterfactual' permet d'approcher la réalité de la situation 'sans'.

information constitue souvent une première et importante source de données pour une évaluation ; par conséquent, il est aussi essentiel d'obtenir un aperçu de la qualité de l'information. En ce qui concerne la fiabilité de la collecte de données (2.1.8), les principales conditions pour une collecte fiable (échantillon suffisamment grand, triangulation, bonne fréquence, indépendance du responsable de la collecte) sont généralement suffisantes pour l'efficacité et l'efficience, mais (beaucoup) moins pour l'impact, la durabilité et la pertinence[38].

*Cadre 3 : Information de base claire quant au fonctionnement du système S&E chez PROTOS (programme au Rwanda)*

L'ONG Protos a élaboré un document 'Scénario pour le suivi des indicateurs' dans lequel il est clairement indiqué, par indicateur, quelles personnes et organisations sont impliquées dans la collecte de données. Il est également spécifié qui analysera les données. En outre, ces scénarios précisent par quelle méthode, quand et avec quelle périodicité les données seront collectées pour l'indicateur en question. Il est aussi mentionné quelles formations sont éventuellement prévues par rapport à une méthode ou un instrument spécifique. De tels scénarios de suivi peuvent être utiles, avant tout, pour garder une vue d'ensemble lorsque différents partenaires sont chargés de parties spécifiques de la collecte et de l'analyse des données, mais ils peuvent aussi être importants pour donner à un évaluateur externe un aperçu des responsabilités et des procédures de collecte et d'analyse des données.

### 3.2.2 Le système S&E dans la pratique

Sous ce composant, on examine comment le système S&E fonctionne (ou a fonctionné) dans la pratique et 12 items ont été définis, lesquels correspondent en partie à ceux de la section 1.3 (*Le système S&E proposé*). La qualité du système S&E proposé, alliée à la qualité de l'application de ce système, est probablement le composant qui influence le plus l'évaluabilité d'une intervention. Les faiblesses dans d'autres composants peuvent, dans une large mesure, être compensées par un système S&E bien élaboré et qui fonctionne bien. À l'inverse, l'évaluabilité pratique d'une intervention sera faiblement cotée si la fonction S&E est défaillante, même si l'on dispose d'un plan d'intervention de bonne qualité.

**Tableau 9: Le système S&E dans la pratique**

| | Pertinence | Efficacité | Efficience | Impact | Durabilité | Index d'évaluabilité |
|---|---|---|---|---|---|---|
| 2.2 Le système S&E (dans la pratique) | 3,03 | 3,30 | 3,50 | 2,53 | 2,85 | 3,04 |
| 2.2.1 La vision sur et le rôle du suivi et des évaluations (S&E et évaluations indépendantes) sont claires | | 3,23 | 3,31 | 2,90 | 2,73 | 3,05 |
| 2.2.2 Suffisamment de temps, ressources et personnel ont été prévus pour un fonctionnement adéquat du système S&E | 3,67 | 3,87 | 4,15 | 3,21 | 3,63 | 3,71 |
| 2.2.3 Les responsabilités et procédures quant à la collecte et l'analyse des données S&E sont claires | 3,50 | 3,85 | 4,08 | 2,64 | 3,41 | 3,50 |
| 2.2.4 Les responsabilités et procédures quant à la prise de décision sur base de l'analyse des données S&E sont claires | 3,62 | 3,75 | 4,03 | 2,95 | 3,56 | 3,58 |
| 2.2.4bis Les parties | 4,18 | 4,20 | 4,33 | 4,23 | 4,18 | 4,22 |

[38] Cet item n'a pu être coté que pour 24 interventions.

| | Pertinence | Efficacité | Efficience | Impact | Durabilité | Index d'évaluabilité |
|---|---|---|---|---|---|---|
| concernées les plus importantes sont d'accord avec le système S&E proposé (y inclus leur rôle dans ce système) [39] | | | | | | |
| 2.2.5 Le personnel en charge du S&E est compétent et indépendant | 2,95 | 3,11 | 3,32 | 2,73 | 3,00 | 3,02 |
| 2.2.6 Le système S&E de l'intervention est articulé avec le système S&E national/local | 2,59 | 2,88 | 3,12 | 2,52 | 2,64 | 2,75 |
| 2.2.7 Il existe une motivation interne pour la gestion stratégique et l'apprentissage | 3,72 | 3,80 | 3,90 | 3,70 | 3,72 | 3,77 |
| 2.2.8 Les résultats du S&E sont effectivement utilisés pour l'apprentissage | 3,05 | 3,08 | 3,30 | 2,93 | 3,00 | 3,07 |
| 2.2.9 Les résultats du S&E sont effectivement utilisés pour rendre compte (pour assurer la redevabilité) | 3,23 | 3,40 | 3,63 | 2,95 | 3,08 | 3,26 |
| 2.2.10 Des évaluations et/ou études ont été faites qui sont de bonne qualité et ont fourni des informations utiles[40] | 2,67 | 2,78 | 2,75 | 2,22 | 2,58 | 2,61 |
| 2.2.11 La qualité du système S&E est régulièrement revue et le système est éventuellement adapté | 2,74 | 2,90 | 3,00 | 2,80 | 2,85 | 2,86 |

Pour *tous* les critères, la pratique S&E *effective* obtient des scores sensiblement (plus de 20%) plus élevés que le système S&E *proposé* dans le plan d'intervention (voir le tableau 3 et le point 1.3 de la partie 1 du présent chapitre). Ceci confirme l'hypothèse formulée précédemment selon laquelle on accorde relativement peu d'attention à la présentation du système S&E dans la proposition d'intervention, pour différentes raisons (voir 1.3). Mais la grande différence dans les scores est aussi une illustration de l'importance qui est accordée au S&E dans la *mise en œuvre* de l'intervention, et a des implications positives pour l'évaluabilité. Le S&E, en particulier le suivi, est donc généralement mieux ancré dans la pratique qu'on pourrait le supposer au vu de la proposition d'intervention. C'est évidemment une bonne chose, non seulement du point de vue de l'évaluabilité, mais surtout parce qu'un (bon) S&E constitue une composante importante de la gestion d'intervention et peut grandement contribuer à l'amélioration des interventions.

Par ailleurs, il existe une forte corrélation entre les scores relatifs à la qualité du plan de S&E (tel que présenté dans la proposition de projet) et la qualité de la pratique S&E, et cela pour tous les critères d'évaluation à l'exception de l'impact[41]. La corrélation est la plus nette pour l'efficacité et la durabilité, ce qui donne à penser qu'il est important de

[39] Cet item faisait initialement partie de la dimension 1 et a été intégré plus tard dans le composant 2.2. Pour des raisons pratiques le numérotage initial a été maintenu.

[40] Les données pour cet item portent seulement sur 27 des 40 interventions analysées. Dans les 13 autres interventions, nous ne pouvons pas attendre qu'une évaluation ou étude aurait eu lieu déjà.

[41] Ce dernier point est probablement lié au fait que l'impact fait l'objet de peu d'attention explicite dans les systèmes S&E (tant dans la proposition que dans la pratique), ce qui réduit aussi la possibilité d'avoir une concordance entre la proposition et la pratique.

bien intégrer – entre autres – ces critères dès le début dans le système S&E. Malgré le fait qu'il existe une forte corrélation, on a trouvé, parmi les 40 interventions étudiées, des exemples dans lesquels la qualité médiocre du système S&E dans la proposition d'intervention a été corrigée par la suite, par exemple en s'engageant résolument dans le développement d'un système S&E au cours du premier semestre de la mise en œuvre de l'intervention. L'avantage d'une telle approche est que l'appropriation (*ownership*) du système peut être mieux assurée et qu'il est plus simple d'impliquer d'autres acteurs dans l'élaboration et l'application du système. Par ailleurs, une forte corrélation n'implique pas forcément la causalité (du papier vers la pratique) : comme nous l'expliquons ci-après, il est possible que la pratique S&E influence aussi la proposition S&E sur papier, ce qui peut être le cas, entre autres, lors d'interventions qui font suite à des interventions antérieures.

D'autre part, ce composant n'enregistre que des scores moyens si l'on le compare aux sept autres composants analysés (tableau 3). Ceci indique qu'il reste encore une marge d'amélioration. Ce qui nous semble normal, car comme les entretiens avec les exécutants des interventions nous l'ont appris, dans de nombreux cas, les systèmes S&E existants avaient été introduits ou mis en œuvre récemment. Dans bon nombre d'interventions, du reste, le processus n'est pas encore achevé, mais l'intention de parachever tant la structure que l'application du système est présente. Ce travail de parachèvement consiste avant tout à forger un véritable système à partir des différents composants qui, la plupart du temps, sont déjà présents (au moins à l'état d'ébauche), mais ne sont pas encore accordés entre eux. Un autre constat est que dans de nombreuses interventions, les 'bonnes pratiques' en matière de S&E semblent être présentes, mais ne sont documentées d'aucune manière et sont aussi, bien souvent, dissociées des autres pratiques S&E. Du point de vue de l'évaluabilité, on peut donc supposer que celle-ci augmentera encore dans le futur à mesure que la pratique S&E se développera.

Il semble par ailleurs que l'on arrive déjà à assurer un suivi correct sur le plan opérationnel : dans bien des cas, l'utilisation des moyens, le suivi des activités et la progression au niveau des outputs font désormais partie intrinsèque de la pratique de mise en œuvre. Toutefois, la composante évaluative du système S&E semble recevoir moins d'attention, en particulier au niveau de l'impact et de la durabilité. Les causes possibles de ce manque d'attention sont analysées ci-après.

Si nous comparons les scores des cinq critères CAD, un schéma comparable à celui des autres composants se dessine, à savoir que le score le plus faible se situe au niveau de l'impact et le plus élevé au niveau de l'efficience. On observe en outre que les scores présentent le même ordre que pour la description du système S&E dans la proposition d'intervention: cette description, avec toutes ses imperfections, est donc une préfiguration de la pratique ultérieure. Les faibles scores pour l'impact, mais aussi pour la durabilité, impliquent que les défis qui sont déjà présents intrinsèquement (ex ante) pour ce qui est de l'inclusion de ces deux critères dans les évaluations, sont encore amplifiés. De l'autre côté, l'efficacité et surtout l'efficience obtiennent des scores relativement bons, ce qui est positif du point de vue de l'évaluabilité, d'autant que ces deux critères, dans bon nombre d'évaluations, occupent une place centrale.

Il ressort de notre analyse que la vision sur et le rôle du suivi et des évaluations au niveau intervention sont relativement peu développés, ce qui est assez surprenant à la lumière de la pratique S&E en cours de développement (2.2.1), même s'il y a une série d'exceptions intéressantes où des changements dans la politique donnent lieu à l'élaboration d'une approche et d'une politique S&E et sont soutenus par des moyens suffisants en personnel (voir le cadre 4). Dans ce cadre, le fait que l'impact et la durabilité affichent des scores plus faibles n'est pas étonnant ; ce qui l'est plus, c'est que la différence est considérable. À l'évidence, ces deux critères sortent du champ de vision lors du développement d'une politique S&E. En creusant un peu plus, nous trouvons encore plusieurs autres raisons. Ainsi, il est fort possible que la politique suive, jusqu'à un certain point, la pratique (et non l'inverse) et que cette pratique soit, à ce jour, correctement mise en œuvre avant tout sur le plan opérationnel. Si tel est le cas, il est

logique que le suivi soit généralement mieux élaboré que l'évaluation dans les documents de politique. Il est clair, par ailleurs, que la définition d'une vision et d'une politique en matière de suivi et d'évaluation est souvent pilotée au départ de l'organisation (belge) responsable, mais qu'il faut du temps pour traduire cette politique en interventions, notamment en raison du fait qu'une 'nouvelle' politique – ou une autre politique – ne peut en principe être appliquée qu'après concertation avec les partenaires. Dans bien des cas, ce processus n'est pas encore achevé, ce qui est compréhensible si l'on considère que l'attention accrue pour le S&E est relativement récente. Tout cela semble avoir comme conséquence, en ce qui concerne l'évaluabilité, que la situation est globalement positive sur le plan de l'efficacité et de l'efficience (et aussi, pour d'autres raisons, de la pertinence), mais que c'est moins le cas pour l'impact et la durabilité.

*Cadre 4 : Comment un changement de politique peut induire une amélioration de l'approche S&E*

Ces dernières années, la stratégie de Médecins Sans Vacances a subi une modification en profondeur. Si autrefois l'accent était mis sur l'envoi de médecins belges qui allaient sur place fournir des prestations médicales, le travail s'articule aujourd'hui autour des besoins des organisations partenaires. Ce changement fondamental se répercute aussi dans l'instrumentation S&E. L'organisation a développé toute une gamme d'instruments S&E qui doivent permettre d'assurer un meilleur suivi des effets du programme. L'élément central est ici un instrument visant à mesurer les réalisations sur le plan du renforcement des capacités, ceci en partant des changements définis/souhaités par le partenaire lui-même et en définissant, entre autres, des '*progress markers*' qui font l'objet d'un suivi. Par ailleurs, la qualité et les effets des missions des médecins belges sont systématiquement mesurés via, entre autres, un rapport de mission rédigé en commun (par le médecin et le partenaire) dans un format standardisé et un rapport spécifique rédigé par chacune des parties. Ces documents sont ensuite réunis dans une note de synthèse. Sur la base de cette synthèse, un feed-back est donné à toutes les parties concernées.

On remarque aussi, dans le tableau 9, le score très élevé quant au niveau d'accord entre les acteurs les plus importants en ce qui concerne leur rôle dans le système S&E (2.2.4bis). Nous devons cependant nuancer quelque peu ce score, car la plupart du temps, seule l'équipe de projet s'occupe du S&E, les autres acteurs clés n'ayant qu'une contribution (attendue ou demandée) très limitée. Il ne semble y avoir des problèmes que dans les interventions ou organisations plus complexes, par exemple lorsque les niveaux supérieurs formulent des exigences excessives (aux yeux des échelons inférieurs) en matière de collecte et traitement des données et de rapports.

L'étude montre aussi que les moyens (temps, personnel, fonds) pour le S&E sont, en général, présents en suffisance, même si le niveau impact, pour des raisons évidentes, enregistre ici un score plus faible (2.2.2). Du point de vue de l'évaluabilité, ceci est un constat positif, qu'il convient toutefois de nuancer quelque peu. Des discussions plus approfondies avec les équipes de projet ont en effet révélé que si un temps suffisant était disponible pour assurer les fonctions S&E de base, comme la collecte et le traitement des données, le temps manquait souvent pour examiner et analyser ces données en détail. Dans la plupart des interventions, en effet, le S&E n'est pas confié spécifiquement à une personne (ou un service), mais la fonction S&E est assurée par plusieurs membres de l'équipe, chacun étant généralement responsable de la collecte des données concernant (une partie de) 'son' output, et reste donc limitée au niveau opérationnel (inputs – activités – outputs). Le focus sur le niveau opérationnel implique donc que les équipes de projet s'occupent essentiellement de l'aspect 'ici et maintenant', et qu'il y a peu de moments prévus dans les processus pour examiner les choses avec une certaine distance. Ceci peut aussi expliquer le score relativement faible en ce qui concerne la révision (régulière) de la qualité des systèmes S&E (2.2.11). Une autre raison pour laquelle les moyens sont jugés insuffisants est l'attention relativement limitée qui est consacrée à l'évaluation (au sens large du terme) ; ceci sera développé plus en détail dans la partie 3 de ce chapitre.

Les responsabilités et les procédures quant à la collecte et l'analyse des données S&E et à la prise de décision sur ce plan sont bien définies, à l'exception du niveau impact, celui-ci n'étant généralement pas pris en compte (2.2.3 et 2.2.4). Bien souvent, le suivi (et l'évaluation) aux niveaux supérieurs ne fait tout bonnement pas partie des tâches

S&E ; une telle analyse est considérée comme une responsabilité collective, si bien que personne ne se sent spécifiquement concerné[42]. Bien qu'il existe une saine "pression des pairs", chacun semble s'occuper "uniquement" du suivi de "sa" partie dans la mise en œuvre de l'intervention. Le suivi des effets *globaux* de l'intervention, et en particulier de l'impact, se situe à un niveau plus élevé et nécessite en outre une approche plus spécifique et très astreignante sur le plan de la collecte et de l'analyse des données, pour laquelle des connaissances méthodologique spécifiques et des moyens plus étendus sont requis. Il n'y a, souvent, pas réellement de place réservée à cet effet dans l'élaboration des systèmes S&E, et il n'y a dès lors pas de responsabilités définies à ce niveau. En outre, l'information qui est obtenue par le biais de l'analyse d'impact dépasse souvent le niveau spécifique d'intervention, si bien qu'au sein des interventions/organisations, elle est jugée moins utile et moins utilisable directement, et n'est dès lors pas une priorité. Pourtant, une telle information est particulièrement utile au niveau (de politique) supérieur, mais présente en quelque sorte un caractère de 'bien public', ce qui explique en partie le manque d'investissement dans les évaluations d'impact et de durabilité par les interventions spécifiques.

Incidemment, nous voudrions faire observer que l'opportunité et la possibilité d'impliquer d'autres acteurs dans le S&E ne sont que rarement envisagées de manière approfondie. Ce faisant, on passe à côté de certaines opportunités sur le plan de l'apprentissage, par exemple, mais aussi en ce qui concerne la pérennisation des systèmes S&E et la possibilité de mieux répartir la charge et la responsabilité du S&E entre les différents acteurs.

*Cadre 5 : Le développement d'une approche S&E au-delà des interventions (CTB Katanga)*

Le DTF du projet EDUKAT prévoit des indications claires pour la mise sur pied et l'organisation du S&E qui se sont inspirées de la politique générale et des guides développés par la CTB dans ce domaine. Dans ce cadre global, l'EDUKAT a pris une initiative originale par la rédaction d'un document 'Plan de travail baseline' qui essaye d'inclure plusieurs activités S&E dans une démarche cohérente. Ce plan de travail fut formulé après une révision du cadre logique initial (considéré comme assez compliqué et pas entièrement cohérent) et définit les étapes d'une démarche d'élaboration d'un baseline devant intégrer la formulation d'une matrice de suivi (dérivée du cadre logique revu), un plan de gestion des risques et l'application effective des mécanismes de suivi. La mise en œuvre de ce plan constitue la responsabilité d'un gestionnaire de programme chargé de la gestion transversale du S&E des différents projets bilatéraux mis en œuvre dans la province de Katanga. Ce choix s'est inspiré de la nécessité d'assurer l'apprentissage et la continuité des systèmes de S&E. Il s'inscrit dans le temps (un engagement de 10-12 ans est prévu) et doit également appuyer un processus de renforcement institutionnel dans la mesure où des structures locales seront associées au processus de collecte de données et d'analyse et pilotage des résultats S&E.

La compétence et l'indépendance du personnel qui est en charge du S&E obtiennent des scores moins élevés que la plupart des autres items dans ce composant (2.2.5). Une première explication réside dans le fait qu'il est difficile, dans les petites interventions, d'organiser un S&E réellement indépendant, par exemple en le confiant à un service ou une personne distinct. Ceci n'est réalisable que dans les interventions ou programmes de plus grande ampleur (voir le cadre 5). En ce qui concerne la compétence, il y a peu d'indications quant à une expertise spécifique ou des investissements dans une formation spécifique. La plupart des responsables qui s'occupent du S&E sont donc formés « sur le tas ». Globalement, toutefois, cela engendre peu de problèmes car (1) il y a beaucoup d'échanges et une bonne culture d'apprentissage parmi les équipes (voir ci-après), (2) la tâche S&E se limite le plus souvent au niveau opérationnel, où il y a moins de difficultés sur le plan méthodologique, (3) de nombreuses interventions ont développé une bonne instrumentation standard pour les activités importantes (p. ex. formations). Il est clair cependant qu'aux niveaux supérieurs, le S&E (et en particulier l'évaluation) pose des exigences plus pointues auxquelles un personnel d'intervention non qualifié ne peut pas toujours répondre. Nous pouvons constater à cet égard que ces

[42] Dans certains cas, les acteurs locaux doivent aussi satisfaire à des exigences de suivi qui sont imposées par différents acteurs et qui souvent sont mal accordées entre elles.

difficultés méthodologiques ne sont pas suffisamment identifiées à ces niveaux et que, de facto, le niveau outcome et impact fait l'objet de peu d'attention dans le cadre du suivi. Un dernier élément qui se dégage est que dans bien des cas, la qualité et la continuité des systèmes S&E souffrent de la rotation du personnel ou des déplacements internes. Ceci est source de problèmes en particulier lorsque le système S&E est encore peu formalisé, ce qui est une nouvelle illustration de l'importance d'une description plus élaborée du système S&E.

Les résultats de notre analyse en ce qui concerne l'utilisation du S&E pour l'apprentissage sont plutôt bons (2.2.7), excepté pour l'impact. Le S&E étant fortement orienté vers le niveau opérationnel, il est logique que le score soit bon sur le plan de l'efficience. L'utilisation des résultats S&E pour assurer la redevabilité (2.2.9) affiche un score légèrement plus élevé que leur utilisation pour l'apprentissage. Ceci peut sans doute s'expliquer, entre autres, par le fait que pour la redevabilité, les données collectées en tant que telles suffisent en grande partie, alors que l'apprentissage nécessite une analyse, une réflexion et une appropriation plus poussées et que, dans bien des cas, le temps manque à cet effet, comme nous l'avons déjà souligné. D'un autre côté, la redevabilité (‘accountability’) est souvent interprétée dans un sens restreint et se limite de facto à la redevabilité à l'égard du bailleur de fonds et indirectement (dans certains cas) de l'arrière-ban plus étendu, p. ex. via les sites web. Il n'est que très rarement question d'une redevabilité délibérée à l'égard des acteurs clés, comme les groupes cibles. Un autre élément en lien avec ce qui précède est qu'il y a relativement peu de feedback qui est donné (sur la collecte de données et les résultats sur la base des données analysées) vis-à-vis des acteurs clés locaux (ainsi que, bien souvent, du groupe cible) qui sont impliqués dans les processus de collecte de données. Un tel feedback est essentiel si l'on veut éviter que la collecte de données locale devienne un processus rituel qui influe négativement sur la fiabilité des données (et donc aussi sur l'évaluabilité) et à plus long terme, sur la pérennité du système S&E. Ce besoin de rétroaction vaut non seulement pour la relation entre les responsables locaux des interventions et les acteurs clés (groupes cibles compris), mais aussi pour la relation entre l'organisation belge et les responsables locaux des interventions et celle de la DGD vis-à-vis des acteurs indirects et de la CTB.

Dans bon nombre d'interventions, il y a toutefois une forte motivation pour la gestion stratégique et l'apprentissage, qui se traduit aussi dans la pratique de mise en œuvre (2.2.7). Ceci crée un cadre dans lequel les imperfections éventuelles du système S&E (p. ex. sur le plan des compétences et de l'indépendance et l'attention réduite pour les niveaux supérieurs dans la chaîne moyens-fin) sont malgré tout compensées en partie. Ceci favorise aussi une culture d’intervention ouverte à la critique et au questionnement, ce qui constitue un avantage 'transversal' important pour l'évaluabilité.

Comme nous l'avons souligné précédemment, la pratique d'évaluation dans les interventions reçoit (trop) peu d'attention. Ceci est encore illustré par le faible score – cet item affiche le score le plus faible dans ce composant, et ce pour tous les critères, et figure aussi, pour 3 critères, dans la liste des items avec les scores les plus faibles – en ce qui concerne le 'track record' des interventions étudiées sur ce plan (2.2.10)[43]. Nous y voyons plusieurs explications qui seront abordées plus en détail dans la partie 3. Pour l'heure, nous nous limiterons à mentionner la focalisation sur le suivi dans bon nombre d'interventions, de sorte que l'on considère souvent – à tort ou à raison – que l'évaluation (externe) n'offre aucune plus-value. D'un autre côté, il y a aussi des exemples de bonne pratique, où la mise en place et l'exécution d'évaluations sont déjà, dans une large mesure, définies ex ante et sont appliquées sur une longue période, si bien que ces évaluations peuvent apporter une plus-value appréciable (voir cadre 6).

[43] …en précisant toutefois que nous n'avons pu donner un score à cet item que pour 27 interventions.

*Cadre 6 : Élaboration d'une politique et d'une pratique d'évaluation claires*

En tant qu'ANG active dans le domaine de l'éducation au développement, ITECO a développé une politique claire de planning, monitoring et évaluation. Au niveau de l'évaluation de ses actions de formation, ITECO a développé une démarche type qui se décline en trois dimensions complémentaires couvrant plusieurs maillons de la théorie de changement sous-jacente : l'évaluation des apprentissages acquis lors d'une formation, la mise en œuvre (par les participants) de ces apprentissages et les effets de la formation (à travers la mise en œuvre de leurs acquis) au niveau de l'organisation et de l'environnement des participants. Les résultats de cette analyse permettent à ITECO de dépasser le champ pédagogique pour se former une idée sur sa contribution aux transformations souhaitées.
Au niveau de son programme cofinancé par le gouvernement belge, ITECO a sollicité l'appui d'un spécialiste de l'éducation et de la formation pour peaufiner ses outils de S&E, notamment la définition d'indicateurs adéquats. L'aboutissement de ce processus s'est concrétisé, entre autres, dans un cadre logique de qualité excellente, avec une baseline claire et des cibles annuelles permettant un suivi adéquat du progrès réalisé.

Un dernier point concerne l'articulation des systèmes S&E des interventions avec les systèmes nationaux et locaux (2.2.6). Les scores – et ce n'est sans doute pas surprenant – sont ici assez faibles, sauf pour les interventions au Rwanda (voir chapitre 4.1). Dans de nombreuses interventions, principalement en dehors du canal bilatéral, il s'agit d'une considération dont on ne se préoccupe pas vraiment. Dans les interventions bilatérales, d'une manière générale, il en va autrement, mais ici des difficultés pratiques interviennent souvent, comme la faiblesse ou l'absence de systèmes S&E au niveau national ou local.

Du reste, aucun projet ne semble avoir pour ambition d'améliorer et de pérenniser les systèmes S&E existants. Avoir de telles attentes à l'égard des interventions, c'est sans doute placer la barre trop haut. Néanmoins, chaque projet devrait vérifier explicitement si ses systèmes S&E (ou quelques-uns des indicateurs clés) ne peuvent pas être alignés avec les systèmes nationaux, locaux ou partenaires. Car il est clair que l'absence de lien avec les systèmes nationaux ou locaux a d'importantes implications en ce qui concerne l'évaluabilité, notamment si l'on veut pouvoir évaluer ex post et déterminer, par exemple, la durabilité des bénéfices réalisés au terme d'un projet.

## 3.3 Le contexte de l'évaluation

Pour déterminer l'influence du contexte sur l'évaluabilité des interventions, nous avons pris comme hypothèse que pour chaque intervention, une évaluation externe indépendante devrait être effectuée. Par ailleurs, nous avons choisi de ne pas inclure dans l'analyse certains éléments pratiques liés au contexte, comme la situation sur le plan de la sécurité, de l'infrastructure, etc., car ils sont très spécifiques à chaque intervention et difficiles à manier dans le cadre de cette étude générale. Il doit être clair, cependant, que dans la réalité, les éléments de cet ordre sont d'une grande importance pour déterminer l'évaluabilité, raison pour laquelle il a été décidé de les mentionner explicitement dans le cadre d'analyse.

### 3.3.1 L'attitude des acteurs clés

L'attitude des acteurs clés (notamment les acteurs étroitement impliqués dans la mise en œuvre de l'intervention, mais aussi les instances publiques, le(s) bailleur(s) de fonds et autres organisations actives dans la région) revêt une grande importance pour l'évaluabilité des interventions. Une attitude négative de la part des acteurs rend le travail des évaluateurs très difficile en toutes circonstances, même si les autres conditions – comme un bon plan d'intervention ou un système S&E performant – sont largement remplies.

**Tableau 10: Résultats quant à l'attitude des acteurs clés**

| | Pertinence | Efficacité | Efficience | Impact | Durabilité | Index d'évaluabilité |
|---|---|---|---|---|---|---|
| 3.1 Attitude des acteurs clés | 3,68 | 3,70 | 3,70 | 3,58 | 3,63 | 3,66 |
| 3.1.1 Les utilisateurs les plus importants de l'évaluation ainsi que leurs attentes/intérêts par rapport à l'évaluation ont été clairement définis[44]. | 2,36 | 2,36 | 2,36 | 2,27 | 2,27 | 2,33 |
| 3.1.2 Les acteurs clés sont partie requérante d'une évaluation (ou au moins intéressées) | 3,45 | 3,45 | 3,45 | 3,40 | 3,45 | 3,44 |
| 3.1.3 Les attentes des acteurs clés par rapport à l'évaluation (processus et résultats) sont mutuelle-ment compatibles | 4,11 | 4,11 | 4,19 | 4,11 | 4,16 | 4,14 |
| 3.1.4 Les attentes des acteurs clés par rapport à l'évaluation sont réalistes (compte tenu des moyens disponibles) | 4,32 | 4,23 | 4,49 | 4,11 | 4,32 | 4,29 |
| 3.1.5 Les utilisateurs les plus importants sont associés/seront associés au processus d'évaluation | 3,53 | 3,52 | 3,58 | 3,55 | 3,55 | 3,55 |
| 3.1.6 Les parties concer-nées les plus importantes (y inclus les groupes | 3,62 | 3,62 | 3,62 | 3,56 | 3,56 | 3,59 |

[44] Cet item n'a ou être analysé utilement que pour 22 interventions seulement; les interventions non incluses étaient surtout des interventions qui ont démarré trop récemment pour justifier une évaluation externe.

| | Pertinence | Efficacité | Efficience | Impact | Durabilité | Index d'évaluabilité |
|---|---|---|---|---|---|---|
| cibles) sont associées/ [45]seront associées au processus d'évaluation | | | | | | |
| 3.1.7 Les relations parmi les acteurs clés sont "saines" | 4,26 | 4,24 | 4,24 | 4,22 | 4,22 | 4,24 |
| 3.1.8 Il est possible de contacter les acteurs clés sans que le risque existe qu'ils s'influencent mutuellement | 4,63 | 4,64 | 4,64 | 4,63 | 4,63 | 4,64 |
| 3.1.9 Tous les acteurs clés ont une attitude positive par rapport à l'évaluation indépendante | 4,18 | 4,18 | 4,18 | 4,18 | 4,21 | 4,18 |

Comme il ressort du tableau 10, les scores positifs prédominent. Il s'agit dès lors d'un composant qui affiche globalement un score élevé, avec cinq items dans la liste des 20% d'items présentant les scores les plus élevés (voir aussi le tableau A3 2 de l'annexe 7)[46]. On observe également que les différences entre les cinq critères d'évaluation sont minimes (avec néanmoins, une fois de plus, le score le plus faible pour l'impact et la durabilité), ce qui peut s'expliquer par le fait que l'attitude des acteurs clés concerne le plus souvent 'l'évaluation indépendante' *dans son ensemble*. Il est clair qu'un bon score global en ce qui concerne l'attitude des acteurs clés est un constat positif dans l'optique de l'évaluabilité. Une telle attitude positive joue un rôle non seulement pendant le processus d'évaluation, mais est aussi, bien souvent, le signe d'une ouverture au questionnement et à la critique, un autre facteur important pour une bonne mise en œuvre de l'intervention.

'Les utilisateurs les plus importants de l'évaluation ainsi que leurs attentes/intérêts par rapport à l'évaluation ont été clairement définis' (3.1.1) est l'item qui présente le score le plus faible. Nous voyons à cela plusieurs explications. La première est que – comme indiqué plus haut – la politique et la pratique concernant l'évaluation (indépendante) au niveau intervention sont relativement peu développées (contrairement au suivi). En réalité, on s'attache peu (de manière explicite) à l'évaluation au niveau intervention et, par voie de conséquence, on ne réfléchit guère aux différents paramètres concernant l'évaluation : on s'occupe peu de mettre en balance les objectifs éventuels de l'évaluation (redevabilité – apprentissage – soutien de la politique), et donc aussi les choix qui en découlent quant aux principaux critères d'évaluation et aux éventuels (futurs) utilisateurs des résultats de l'évaluation (qui ne correspondent pas forcément aux acteurs clés lors de la mise en œuvre de l'intervention[47]). Il n'y a que dans la préparation même à l'évaluation que ces points font l'objet de quelque attention, mais ils restent en général trop peu élaborés, comme le montre notamment l'analyse des termes de référence pour les évaluation.
L'absence d'un effort clair (et de choix clairs) en ce qui concerne cet item a une influence sur l'évaluabilité : si les attentes/intérêts des utilisateurs les plus importants ne sont pas correctement identifiés et/ou si des choix clairs n'ont pas été opérés quant aux objectifs de l'évaluation, cela se traduit inévitablement par une évaluabilité moindre et – surtout – une moindre qualité des produits et des effets de l'évaluation. Les nombreux rapports d'évaluation qui reconnaissent eux-mêmes qu'ils n'ont pas pu analyser certains aspects suffisamment en profondeur en sont clairement l'illustration.

[45] Cet item a été analysé au niveau de 26 interventions seulement.

[46] Comme nous le verrons dans la suite du rapport, les résultats peuvent cependant varier d'un pays à l'autre ; voir la partie 4 (analyse comparative) ci-après.

[47] Le groupe des utilisateurs (éventuels) des évaluations peut être différent du groupe des acteurs clés. Ce dernier concerne les acteurs qui ont un intérêt direct dans le projet et sont étroitement impliqués dans la mise en œuvre. Les 'utilisateurs' de l'évaluation comprennent ces acteurs clés, mais peuvent aussi englober d'autres parties comme les bailleurs de fonds, les membres d'un réseau dont le projet fait partie, les services nationaux de planification, …

Le score pour l'item portant sur le fait de savoir si les acteurs clés sont demandeurs d'une – ou au moins intéressés par une – évaluation (3.1.2) est relativement bon, mais se situe malgré tout sous le score moyen des items qui constituent ce composant. Néanmoins, dans 85% des interventions, une majorité au moins des acteurs sont intéressés ou demandeurs. La volonté de procéder à une évaluation externe et/ou l'intérêt pour une telle évaluation peut s'expliquer, entre autres, par l'approche fortement axée sur le résultat dans certaines organisations (p. ex. CTB) ou certains pays (p. ex. Rwanda), mais aussi par la culture d'organisation ouverte et démocratique (particulièrement tangible parmi les ANG). D'un autre côté, une attitude négative n'implique pas automatiquement une aversion pour les évaluations externes. Les interventions qui adoptent cette attitude le font souvent parce qu'elles estiment que les évaluations externes n'apportent pas de plus-value (par rapport aux évaluations internes et aux résultats du suivi) ; dans plusieurs cas, des expériences négatives avec des évaluations externes antérieures ont aussi joué un rôle.

Les scores élevés en ce qui concerne la compatibilité mutuelle entre les attentes des acteurs clés par rapport au processus et aux résultats de l'évaluation et le caractère réaliste de ces attentes (3.1.3 et 3.1.4) méritent une analyse plus nuancée qui montre que leur influence positive sur l'évaluabilité est moins grande qu'on pourrait le croire à première vue. Tout d'abord, on peut observer que les attentes des acteurs clés sont réalistes avant tout parce que le niveau d'ambition par rapport aux évaluations externes n'est généralement pas très élevé. Peu d'interventions, par exemple, veulent réellement investir dans une évaluation approfondie centrée sur l'impact ou la durabilité. D'un autre côté, si on en demande beaucoup aux évaluateurs (en comparaison avec les moyens disponibles), on accepte aussi implicitement qu'il y ait un compromis par rapport à la profondeur de l'analyse. En ce qui concerne la compatibilité mutuelle des attentes, il y a peu de problèmes dans la mesure où la définition des utilisateurs les plus importants et de leurs attentes n'intervient pas systématiquement, comme nous l'avons indiqué ci-avant. De cette manière, les différences éventuelles dans les attentes et les intérêts ne sont pas mises en lumière ou sont dissimulées en formulant des objectifs et des thèmes d'évaluation larges.

L'implication des utilisateurs dans le processus d'évaluation (3.1.5) obtient des scores globalement bons, mais pas très bons. Un élément important est que les cas où les évaluations externes sont monopolisées par un seul acteur sont exceptionnels. L'expérience passée nous a appris que dans des cas pareils, les résultats de l'évaluation sont en général peu exploités. Par contre, une implication adéquate des utilisateurs pose des exigences élevées quant à la mise en œuvre de l'évaluation et à l'évaluabilité. C'est pourquoi il est important de doser correctement cette implication (phase préliminaire, mise en œuvre, phase ultérieure), ce qui signifie être attentif à la faisabilité de l'évaluation afin de ne pas alourdir inutilement le processus. Dans bon nombre d'interventions, cet exercice n'est pas encore effectué (notamment en raison du fait que les utilisateurs et leurs attentes et les objectifs de l'évaluation ne sont pas définis de façon explicite) : la plupart des interventions optent de manière assez intuitive pour une certaine forme d'implication, notamment dans la phase préliminaire et ultérieure, sans vérifier si ce choix est le meilleur.
L'implication des *acteurs clés* (3.1.6) affiche un score légèrement supérieur par rapport aux utilisateurs les plus importants, ce qui s'explique aisément si l'on considère qu'il est plus évident d'impliquer ces acteurs dans l'évaluation. Un constat important à cet égard est qu'à peine un peu plus d'un tiers des interventions impliquent de manière explicite les groupes cibles dans les évaluations. Quelques cas de bonnes pratiques montrent qu'une implication adéquate des utilisateurs et des acteurs clés peut se mettre en place plus facilement dans le cadre d'une approche élargie qui accorde une place centrale à cette implication sur une période plus longue, pas uniquement dans le cadre des évaluations, mais à travers une approche systémique qui commence à la préparation du projet et se poursuit pendant la mise en œuvre (voir cadre 7).

*Cadre 7 : Une approche large des évaluations assure une forte implication des acteurs clés*

Le projet PARZS de la CTB au Benin (*Projet d'Appui au Renforcement des Zones et Départements Sanitaires du Mono-Couffo et de l'Atacora-Donga*) dans le secteur de la santé a investi beaucoup dans l'application des principes d'outcome mapping (OM) dans son système de planification et de suivi et évaluation. La définition des changements de comportement et des marqueurs de progrès par les différents acteurs ou instances impliqués dans le projet sont à la base de ses systèmes de suivi et évaluation. Dans le système S&E du projet, plusieurs moments d'auto-évaluation et de contrôle sont prévus qui permettent un bon suivi. Des évaluations par des pairs et le système de OM sont adaptés au contexte local et prennent en compte les capacités (de suivi) et les attentes des parties prenantes. Au cours du projet, le système a graduellement pris forme et est utilisé par les différentes parties prenantes qui sont responsabilisées pour ce travail. Plusieurs principes et éléments du système sont repris dans le S&E du secteur de santé qui est basé sur les principes de FBR (Financement Basé sur les Résultats).

Dans la plupart des interventions, les relations entre les acteurs clés sont saines (3.1.7). Lorsque ce n'est pas le cas, c'est souvent lié à des incidents survenus dans le passé ou à des oppositions entre acteurs qui existent de longue date et qui, parfois, dépassent le contexte spécifique. Dans la plupart des interventions, les relations entre les acteurs sont sainement critiques, ce qui implique qu'ils ne sont pas excessivement bienveillants entre eux. Il est possible, toutefois, qu'une définition plus 'pointue' des intérêts des acteurs clés et des objectifs et du contenu de l'évaluation puisse rendre plus pointues les relations, mais cela ne doit pas nécessairement être négatif.

Le fait que les relations entre les acteurs clés sont généralement saines est assurément un facteur qui explique le score élevé (le plus élevé pour ce composant) attribué à l'item '*Il est possible de contacter les acteurs clés dans que le risque existe qu'ils s'influencent mutuellement*' (3.1.8). Par ailleurs, il est clair également que dans la plupart des interventions, il y a une bonne compréhension du rôle de l'évaluation (externe) dans le processus de développement et de la manière dont on doit se positionner par rapport à un tel processus. Ceci ressort également du score élevé obtenu par l'item portant sur l'attitude des acteurs clés par rapport à l'évaluation indépendante (3.1.9). Dans plus de 80% des interventions, tous les acteurs clés ou presque ont une attitude positive par rapport aux évaluations indépendantes, ce qui nous semble une donnée importante dans l'optique de l'évaluabilité.

### 3.3.2 Le contexte plus large

Ce dernier composant de notre cadre d'analyse se penche sur le contexte plus large, autrement dit le contexte au-delà de l'environnement direct des interventions. Il est clair que ce contexte peut influencer l'évaluabilité, dans un sens positif comme négatif.

Parmi tous les composants, le contexte plus large est celui qui obtient les meilleurs scores (les trois items de ce composant font partie des 20% d'items avec les meilleurs scores - voir aussi le tableau A3 2 de l'annexe 7), malgré le fait que dans l'étude ont été repris des pays dont le cadre institutionnel et politique peut poser certains défis en ce qui concerne les évaluations externes (tableau 3). Il y a peu de différences entre les différents critères et comme on pouvait s'y attendre, l'impact affiche un score légèrement inférieur. C'est aussi le cas pour la pertinence, ce qui est lié au fait que pour l'évaluation de la pertinence, les difficultés dans le contexte plus large ont plus de poids que pour les autres critères, qui se concentrent plutôt sur l'intervention en tant que telle.

### Tableau 11: Aperçu des résultats quant au contexte plus large

| | Pertinence | Efficacité | Efficience | Impact | Durabilité | Index d'évaluabilité |
|---|---|---|---|---|---|---|
| *3.2 Le contexte plus large* | 4,10 | 4,25 | 4,25 | 4,10 | 4,15 | 4,17 |
| 3.2.1 Le contexte institutionnel et politique plus large est positif par rapport à des évaluations indépendante | 4,40 | 4,45 | 4,45 | 4,40 | 4,45 | 4,43 |
| 3.2.2 Le contexte socio-culturel au niveau des groupes cibles rend possible une collecte correcte de l'information | 4,35 | 4,45 | 4,45 | 4,35 | 4,40 | 4,40 |
| 3.2.3 L'expertise locale avec le profil requis est disponible pour l'évaluation | 4,25 | 4,30 | 4,25 | 4,20 | 4,25 | 4,25 |

Quoi qu'il en soit, les scores élevés sont une nouvelle illustration du fait que des évaluations indépendantes peuvent, globalement, être mises en œuvre sans difficultés notables. Les entretiens avec les parties concernées nous ont appris, il est vrai, que des difficultés se présentent dans chaque évaluation ou presque, mais que l'on trouve assez facilement des solutions garantissant une bonne mise en œuvre de l'évaluation, et donc aussi une bonne évaluabilité.

La constatation ci-dessus doit cependant être mise en perspective. Premièrement, il n'est pas à exclure que les facteurs contextuels qui mettent potentiellement l'évaluabilité sous pression aient aussi influencé notre propre étude, p. ex. par des réponses socialement souhaitables entraînant des scores positifs. Plus généralement, les chercheurs ont eu trop peu de temps pour analyser en profondeur l'influence du contexte (p. ex. pour mener des entretiens avec un panel plus large d'utilisateurs (potentiels) de l'évaluation). En réalité, une étude approfondie au niveau des pays serait nécessaire pour pouvoir estimer réellement l'influence du contexte. La présente étude, vu le temps et les moyens limités, a dû se résoudre à se concentrer avant tout, en ce qui concerne le contexte de l'évaluation, sur l'aspect technique de l'évaluation et non l'aspect politique. Avec une interprétation aussi 'étroite' du contexte, il n'est guère étonnant que des solutions soient aisément trouvées par rapport aux difficultés qui se présentent.

Par ailleurs, il est aussi nécessaire, selon nous, de mener une réflexion approfondie sur la question de savoir si l'on n'est pas trop optimiste quant aux possibilités d'évaluer de manière *indépendante*. À cet égard, il est avant tout important de prendre en compte quelques-unes des constatations faites au sujet du composant précédent. Nous avons ainsi constaté (voir 3.1) que la définition des futurs utilisateurs de l'évaluation et de leurs attentes et intérêts enregistrait un score assez faible. Du fait qu'une attention insuffisante est accordée à cet aspect, il y a un risque significatif que les exercices d'évaluation soient organisés et élaborés essentiellement 'à l'intérieur du système'. En effet, dans les évaluations indépendantes également, cette indépendance concerne avant tout la *mise en œuvre* – par des experts indépendants – de l'évaluation, mais ces experts s'occupent beaucoup moins – voire pas du tout – des objectifs et des questions d'évaluation auxquel(le)s l'évaluation devra, en première instance, apporter une réponse, ceci en raison du fait qu'ils ne sont pas impliqués dans l'élaboration des termes de référence. De la sorte, il est possible que même des évaluations 'indépendantes' soient largement orientées et écartent ou relèguent au second plan des sujets controversés (qui pourraient, par exemple, embarrasser certaines parties). Si les évaluations laissent de côté les sujets délicats, il est évident qu'il y aura moins de problèmes sur le plan du contexte institutionnel et politique ; nous pouvons parler ici d'une forme d'« autocensure » qui apparaît aussi fréquemment dans d'autres situations. D'un autre côté, des évaluateurs compétents et habiles peuvent essayer d'attirer aussi l'attention sur certains aspects *'out of the box'*. Quoi qu'il en soit, il est clair que dans un

tel contexte, le potentiel d'apprentissage d'une évaluation, mais aussi sa pertinence globale, sont entre autres sous pression. Ce constat nous incite à ajouter une nuance à la définition de l'évaluabilité dans le sens où « l'évaluabilité » implique aussi que tout ce qui *doit* en principe être évalué *peut* effectivement être évalué.

L'item '*Le contexte socio-culturel au niveau des groupes cibles rend possible une collecte correcte de l'information*' (3.2.2) affiche lui aussi un score élevé. Cela ne veut pas dire qu'au niveau des groupes cibles, il n'y a pas de difficultés à cet égard, mais qu'elles sont (suffisamment ?) identifiées et que l'on dispose (ou que l'on pense disposer) de solutions pour y faire face. De même, en ce qui concerne certains thèmes 'difficiles' comme par exemple la violence liée au genre ces dernières années dans l'est du Congo, une expérience considérable (en matière d'évaluation) a été accumulée, si bien que les difficultés sont moins grandes qu'il y a quelques années. D'un autre côté, il existe sur ce plan un danger identique à celui décrit au paragraphe précédent, à savoir que certains sujets délicats sont écartés ou ne sont pas réellement traités. D'autre part, l'équipe de recherche a l'impression que les défis méthodologiques de l'étude au niveau des groupes cibles sont parfois sous-estimés[48]. Le fait que l'on entretienne de bonnes relations avec les groupes cibles ne justifie pas, en effet, que l'on applique moins strictement les exigences méthodologiques ou autres en matière de bonne recherche. En outre, les interventions sont toujours plus complexes, avec une multitude d'acteurs, et requièrent dès lors une grande expertise en matière d'évaluation. Enfin, on observe dans bien des cas une « fatigue » sur le plan de l'évaluation et de la recherche qui peut générer des résultats largement faussés.

Les cas où les interventions ont du mal à recruter des évaluateurs avec le profil requis (expertise technique et méthodologique suffisante, indépendance, disponibilité à un tarif conforme au marché) sont exceptionnels (3.2.3). Ce score positif est assurément une illustration de la disponibilité croissante d'experts locaux (entre autres), ce qui est très important du point de vue de l'évaluabilité. D'un autre côté, ceci doit à nouveau être quelque peu nuancé. Les 40 interventions de notre échantillon étant pour la plupart des interventions 'classiques', trouver de bons experts ne pose guère de problèmes. Les expériences avec l'évaluation de formes moins classiques de coopération, comme l'aide budgétaire, montrent que ce n'est pas toujours le cas[49]. Un autre élément qui peut jouer est qu'en l'absence (ce qui est souvent le cas) d'objectifs et de questions d'évaluation clairement définis, les exigences par rapport à l'expertise requise sont aussi définies avec trop peu d'acuité.

### 3.3.3 Éléments pratiques

Pour être complets, nous voudrions rappeler qu'il existe encore un troisième composant qui est important pour l'analyse du contexte de l'évaluation et pour déterminer l'évaluabilité. Il s'agit d'une série d'éléments pratiques ayant (pour la plupart) une influence directe sur l'évaluabilité, comme la situation sécuritaire, les conditions climatiques, l'état des infrastructures locales, l'étendue géographique de la zone d'intervention et l'accessibilité, etc. Ensuite, le timing de l'évaluation peut aussi influencer fortement l'évaluabilité : il est préférable, par exemple, que le travail de terrain pour les évaluations ne coïncide pas avec des élections, des fêtes religieuses, une période de vacances, des périodes fort chargées dans le cadre de l'intervention, etc. De même, certains événements (p. ex. conflits au sein des organisations ou interventions) ou des initiatives similaires (p. ex. d'autres donateurs) peuvent créer des complications. Ces éléments étant, par nature, très spécifiques à l'intervention, nous avons choisi de ne pas les inclure dans notre analyse.

---

[48] Nous n'avons pas pu étudier suffisamment cet aspect – important – pour pouvoir nous prononcer de façon plus catégorique.

[49] Dans le cahier des charges, on avait déjà choisi de ne pas inclure l'aide budgétaire dans cette étude.

# 4 Analyse comparative

Ce chapitre s'appuie sur l'analyse réalisée au chapitre 3 pour tenter de déterminer dans quelle mesure il y a des différences dans l'évaluabilité sur la base de certains paramètres qui avaient déjà été appliqués auparavant pour la détermination de l'échantillon[50]. Il s'agit plus précisément (1) des quatre **pays**, chacun étant représenté dans l'échantillon avec 10 interventions, (2) du **degré de complexité** des interventions, où nous avons établi une distinction entre les interventions avec une TdC 'complexe' et 'moins complexe', 40% des interventions ayant été cataloguées comme 'moins complexes' et 60% comme 'complexes'[51], et (3) du type d'**acteur**, 25% des interventions faisant partie de la coopération bilatérale, 50% de la coopération via des ONG et des syndicats, et 25% via d'autres acteurs (APEFE/VVOB, coopération universitaire, BIO, IMT, BOS+, VVSG). Cette analyse s'appuie en grande partie sur un traitement statistique des scores commentés plus en détail à l'annexe 7.

## 4.1 Comparaison de l'évaluabilité au niveau pays[52]

Le tableau ci-dessous présente les scores au niveau *composants* et pour ce qui concerne les trois grandes dimensions du cadre d'étude. Il est important, à cet égard, de signaler que des scores similaires au niveau composants peuvent dissimuler des différences entre les pays au niveau *items*. Ces différences potentielles sont analysées en comparant les scores moyens par critère CAD au niveau composants et items entre les quatre pays.

Il ressort du tableau ci-dessous qu'aucun schéma clair ne se dégage *globalement* en ce qui concerne les différences entre les scores d'évaluabilité des quatre pays[53]. Il n'y a une différence claire que là où une telle différence était la plus prévisible, à savoir sous la dimension 3 où l'influence du contexte sur l'évaluabilité a été analysée : la RDC et plus encore le Rwanda affichent ici, à première vue, des scores plus faibles que les deux autres pays. Les différences relativement réduites entre les scores pour les autres dimensions et composants donnent à penser que *d'autres paramètres* que le pays sont peut-être plus importants pour expliquer les différences entre les scores (voir l'analyse aux points 4.2 et 4.3 ci-après).

---

[50] Pour plus d'informations concernant l'échantillonnage, voir le chapitre 2.3.1, l'annexe 4 (liste de toutes les interventions reprises) et l'annexe 2, où des informations complémentaires sont fournies sur la méthodologie de recherche.

[51] Initialement, l'étude visait à établir une comparaison entre les interventions dans les secteurs "hard" et "soft", partant de l'hypothèse selon laquelle les interventions sont sans doute plus évaluables dans les secteurs "hard" que dans les "soft", mais il s'est avéré que ce n'était pas faisable. Après une assez longue réflexion interne, il est apparu que c'était surtout la complexité de la TdC (théorie de changement) de l'intervention qui était importante, et il a été procédé à une distinction entre interventions 'complexes' et 'moins complexes', à savoir les interventions avec une TdC complexe et moins complexe. Voir le point 2.3.1 et le point 2 de l'annexe 2 (Description de l'approche et de la méthodologie de recherche) pour plus de détails.

[52] Cette section est basée entre autres sur les quatre notes de pays qui ont été rédigées et partagées avec les organisations concernées ; ces notes, toutefois, n'ont pas de caractère officiel.

### Tableau 12: Aperçu des scores d'évaluabilité par pays

| | Belgique | Bénin | RDC | Rwanda | Index d'évaluabilité (°) |
|---|---|---|---|---|---|
| **Dimension 1 (plan d'intervention)** | **3,14** | **2,94** | **3,24** | **3,32** | **3,16** |
| 1.1 L'analyse sous-jacente | 3,25 | 3,80 | 3,15 | 4,30 | 3,65 |
| 1.2 La logique d'intervention et la théorie de changement | 2,70 | 2,63 | 3,58 | 2,95 | 2,96 |
| 1.3 Le système S&E proposé | 2,55 | 2,18 | 2,34 | 2,98 | 2,51 |
| 1.4 Consistance et adaptation de la logique d'intervention et la théorie de changement | 4,08 | 3,17 | 3,90 | 3,03 | 3,54 |
| **Dimension 2 (la pratique de mise en oeuvre)** | **2,98** | **2,67** | **3,05** | **3,15** | **2,96** |
| 2.1 Information de base quant à la mise en oeuvre de l'intervention | 2,88 | 2,68 | 2,90 | 3,08 | 2,88 |
| 2.2 Le système S&E dans la pratique | 3,08 | 2,66 | 3,20 | 3,22 | 3,04 |
| **Dimension 3 (contexte)** | **4,17** | **4,18** | **3,85** | **3,45** | **3,91** |
| 3.1 L'attitude des acteurs clés | 3,78 | 3,76 | 3,50 | 3,58 | 3,66 |
| 3.2 Le contexte plus large | 4,56 | 4,60 | 4,20 | 3,32 | 4,17 |
| **Score global évaluabilité (°°)** | ***3,28*** | ***3,08*** | ***3,29*** | ***3,28*** | ***3,23*** |

(°) Le score portant sur l'index d'évaluabilité a été calculé sur base d'indexes (*pas* sur base de moyennes). D'autre part, les scores des pays sont des moyennes (*pas* des indexes). Ainsi il est possible que le score global d'évaluabilité diffère légèrement de la moyenne des quatre scores de pays. Ces différences sont cependant minimales et ne changent rien au contenu de l'analyse.
(°°) Pour le score global de l'évaluabilité, une moyenne pondérée a été calculée avec un poids identique pour les dimensions 1 et 2, tandis que le poids de la dimension 3 constitue seulement la moitié de celui des deux autres dimensions.

L'analyse statistique plus poussée au niveau composant et surtout au niveau item fait cependant ressortir une série de différences significatives entre les pays (commentées ci-après), sans pour autant qu'un schéma clair puisse être identifié (voir le Tableau A2 3 à l'annexe 7 pour plus de détails). Une analyse plus précise du tableau ci-dessus et une comparaison entre les scores moyens qui ont été obtenus par critère CAD au niveau composant et item[54], combinées avec les résultats de l'analyse statistique, nous permettent de mieux cerner les différences entre les pays.

Les résultats de l'analyse consécutive sont conformes aux observations par rapport au tableau 12 : nous trouvons la différence la plus nette au niveau du contexte plus large (3.2), où le score plus faible pour les interventions au Rwanda est frappant pour les cinq critères CAD[55]. Au niveau item, nous trouvons des différences statistiquement significatives uniquement en ce qui concerne le contexte institutionnel et politique (3.2.1)[56]. Les scores indiquent que ce contexte influence de manière négative l'évaluabilité des interventions au Rwanda, car la mise en œuvre des évaluations indépendantes peut être fortement entravée. Toutefois, le contexte socio-culturel au niveau des groupes cibles (3.2.2) et la disponibilité d'une expertise locale (3.2.3) ne sont pas des facteurs qui, lorsqu'on fait une comparaison entre les pays, influencent différemment l'évaluabilité des interventions. On remarquera aussi les scores relativement bons pour la RDC. Ces scores ne présentent pas de différences significatives par rapport aux scores du Bénin et de la Belgique ; autrement dit, malgré les difficultés politiques et institutionnelles dans le pays, il n'y a pas, sur la base de cette

[54] Les scores moyens par pays figurent dans le tableau A2 1 de l'Annexe 7. Les résultats et une courte description des tests statistiques correspondants ont été repris dans cette même annexe.

[55] … mais malgré tout, le score du Rwanda reste globalement élevé (surtout en comparaison avec les autres dimensions). Il est important, toutefois, de pointer ici les observations qui ont été formulées dans le chapitre précédent sur la manière dont le 'contexte' a été abordé dans l'étude (voir l'analyse dans la partie 3).

[56] Voir aussi l'annexe 7 pour plus d'information sur les analyses statistiques qui ont été effectuées.

étude, d'indications selon lesquelles ces facteurs contextuels influencent (plus) négativement l'évaluabilité des interventions dans le pays[57].

Pour les autres différences significatives, il est plus difficile de trouver une interprétation univoque étant donné que ces différences ne sont pas toujours présentes entre les mêmes pays. Ce que l'on constate, par contre, c'est que lorsque les différences sont significatives, le Rwanda est souvent l'un de ceux qui obtiennent le meilleur score. Il semble donc que le contexte au Rwanda influence les interventions à la fois de manière positive (1.1, 1.2, 1.3, 2.1 et 2.2) et négative (3.2 et 1.4). Nous pouvons l'expliquer par le contexte de grande clarté sur le plan de la politique (tant sur papier que dans la pratique), qui permet au Rwanda de se distinguer des autres pays. La culture de la performance dans le pays fait en sorte, notamment, que pour chaque intervention, le bien fondé et la problématique peuvent/doivent être clairement formulés, que le lien entre l'analyse et les objectifs de l'intervention sont/doivent être clairs et univoques, et que les interventions, plus qu'ailleurs, sont (doivent être) attentives à la relation entre l'intervention et la politique du pays. Cette culture poussée de la performance a aussi pour effet que les systèmes S&E, surtout lorsqu'ils sont axés sur la redevabilité ascendante, sont plus développés. La nécessaire mise en concordance avec les systèmes locaux influence également la durabilité des interventions (voir les scores élevés pour le Rwanda sur le critère durabilité sous les composants 1.2 et 1.3).

Comme indiqué précédemment, vu le nombre limité d'observations pour ce composant, nous devons rester prudents quant à l'interprétation du composant en ce qui concerne la consistance et l'adaptation de la logique d'intervention et de la théorie de changement (1.4)[58]. En ce qui concerne l'évaluabilité de l'efficience, de l'impact et de la durabilité, nous ne trouvons pas de différences significatives entre les pays. De même, il n'a pas de différence significative entre les pays pour ce qui est d'indiquer et d'argumenter les changements éventuels (1.4.1). Les scores ne sont différents que pour les critères de pertinence et d'efficacité pour les items relatifs au traitement des changements dans le système S&E (1.4.3) et la disponibilité de l'information sur la vision et les opinions des parties concernées les plus importantes quant aux changements éventuels (1.4.2). Les interventions au Rwanda, en particulier, affichent des scores plus faibles pour ces sous-items. Une culture poussée de la performance peut influencer de différentes manières les scores attribués pour ce composant. Ainsi, le fait de mettre fortement l'accent sur les '*quick wins*' peut avoir comme conséquence que l'on est moins disposé, dans les interventions, à signaler les difficultés, les échecs et les changements qui en résultent dans les interventions. Ceci est étroitement lié à la tendance à l'autocensure, décrite ci-avant. En outre, vu la vitesse à laquelle les réformes (élaborées par les pouvoirs publics) sont effectivement mises en œuvre au Rwanda, il est impossible pour les interventions de rapporter tous les changements que cela occasionne. Il en résulte néanmoins qu'il se crée un fossé entre ce qui est sur papier et la réalité, ce qui influence négativement l'évaluabilité.

Quant au score relativement élevé en RDC en ce qui concerne la qualité de la logique d'intervention et de la théorie de changement, nous n'avons pas d'explication immédiate. Un facteur explicatif pourrait être le fait que 7 interventions sur 10 en RDC

---

[57] Du reste, il est important d'observer que les différences entre les pays peuvent aussi être liées à des différences sur le plan du canal d'intervention ('bilatéral' par opposition à 'autre') et du secteur/type d'intervention ('services' par opposition à – p. ex. – 'lobby et advocacy'). On s'est efforcé de composer par pays un échantillon assez similaire, mais cela n'a réussi qu'en partie compte tenu du nombre limité d'interventions par pays et des nombreux paramètres qu'il a fallu prendre en compte ; en outre, l'échantillon est trop réduit pour vérifier les éventuels effets d'interaction. À cet égard, la situation des interventions en Belgique est assez particulière. De nombreuses interventions en Belgique sont fort différentes de celles dans le Sud, car elles développent avant tout une offre pour leurs groupes cibles, qui sont plus des 'consommateurs' que des acteurs impliqués activement ; en conséquence, elles sont moins accessibles et, en général, moins intéressées par les évaluations. Autres problèmes spécifiques dans certaines de ces interventions : les groupes cibles changeants et la difficulté de retracer avec précision des groupes cibles qui avaient été atteints dans des phases antérieures.

[58] Ce point vaut aussi pour les analyses comparatives selon d'autres paramètres, présentées ci-après, et ne sera plus répété à cette occasion.

constituent la deuxième ou troisième phase d'interventions antérieures, ce qui peut avoir contribué à une amélioration progressive de la logique d'intervention et de la théorie de changement.

## 4.2 Comparaison de l'évaluabilité sur la base de la complexité des interventions

Le tableau ci-après présente les scores au niveau *composants* et pour ce qui concerne les trois grandes dimensions du cadre d'étude. Comme pour l'analyse au niveau pays, il est important de signaler que des scores similaires au niveau composants peuvent dissimuler des différences au niveau *items*. Ces différences potentielles sont analysées en comparant les scores moyens par critère CAD au niveau composants et items entre les interventions complexes et moins complexes.

**Tableau 13: Aperçu des scores d'évaluabilité pour des interventions ayant une TdC complexe et moins complexe**

| | Interventions avec TdC moins complexe | Interventions avec TdC complexe | Index d'éva-luabilité (°) |
|---|---|---|---|
| **Dimension 1 (plan d'intervention)** | **2,87** | **3,30** | **3,16** |
| 1.1 L'analyse sous-jacente | 3,31 | 3,83 | 3,65 |
| 1.2 La logique d'intervention et la théorie de changement | 3,03 | 2,92 | 2,96 |
| 1.3 Le système S&E proposé | 2,11 | 2,78 | 2,51 |
| 1.4 Consistance et adaptation de la logique d'intervention et la théorie de changement | 3,03 | 3,67 | 3,54 |
| **Dimension 2 (la pratique de mise en oeuvre)** | **2,82** | **3,07** | **2,96** |
| 2.1 Information de base quant à la mise en oeuvre de l'intervention | 2,68 | 3,03 | 2,88 |
| 2.2 Le système S&E dans la pratique | 2,95 | 3,10 | 3,04 |
| **Dimension 3 (contexte)** | **3,82** | **3,98** | **3,91** |
| 3.1 L'attitude des acteurs clés | 3,44 | 3,80 | 3,66 |
| 3.2 Le contexte plus large | 4,20 | 4,15 | 4,17 |
| **Score global évaluabilité (°°)** | 3,04 | 3,34 | ***3,23*** |

(°) Le score portant sur l'index d'évaluabilité a été calculé sur base d'indexes (*pas* sur base de moyennes). D'autre part, les scores des interventions avec TdC moins complexe et complexe sont des moyennes (*pas* des indexes). Ainsi il est possible que le score global d'évaluabilité diffère légèrement de la moyenne pondérée des scores des interventions avec TdC moins complexe (40%) et complexe (60%). Ces différences sont cependant minimales et ne changent rien au contenu de l'analyse.
(°°) Pour le score global de l'évaluabilité, une moyenne pondérée a été calculée avec un poids identique pour les dimensions 1 et 2, tandis que le poids de la dimension 3 constitue seulement la moitié de celui des deux autres dimensions.

Il ressort du tableau ci-dessus que les 'interventions complexes' (c.-à-d. les interventions avec une TdC complexe) affichent dans l'ensemble (pour 6 des 8 composants) des scores légèrement supérieurs à ceux des interventions moins complexes. C'est une conclusion surprenante, notre hypothèse de départ étant que les interventions complexes sont plus difficilement évaluables que les moins complexes. Il apparaît que cette hypothèse ne se vérifie que pour le composant 1.2, où nous constatons que le score pour '*la logique d'intervention et la théorie de changement*' est un peu plus élevé que pour les interventions moins complexes, mais la différence est tellement minime que l'on ne peut finalement en tirer aucune conclusion. Si nous pouvons, à ce stade, donner une explication quant au constat qui a été fait au niveau de ce composant, c'est sans doute le fait que dans les interventions complexes, il est moins évident d'identifier les éléments critiques, car il y a moins d'évidence empirique sur laquelle on puisse se baser.

Une explication générale pour le score relativement bon des interventions complexes pourrait être que dans ces interventions, en raison précisément de la complexité de la TdC, une plus grande attention est consacrée, par exemple, à l'analyse sous-jacente et au système S&E. Ceci impliquerait que les acteurs concernés accordent plus d'attention à 'l'apprentissage' car ils ont conscience qu'il y a peu d'évidence empirique disponible et qu'il faut donc consacrer plus d'attention à l'analyse et au système S&E afin de pouvoir opérer plus facilement des ajustements. D'un autre côté, on est conscient, du point de vue de la *redevabilité*, du fait que le bailleur de fonds se montre peut-être plus sceptique vis-à-vis des interventions complexes 'difficiles' dont l'efficacité est plus difficile à démontrer.

Pour mieux cerner les différences sur le plan de l'évaluabilité entre les interventions avec une TdC complexe et avec une TdC moins complexe, nous avons établi une comparaison entre les scores moyens qui ont été obtenus par critère CAD au niveau composants et au niveau items[59]. Dans la comparaison entre interventions avec TdC complexe et moins complexe, nous ne trouvons des scores avec des différences significatives que pour 2 composants, à savoir le système S&E proposé (1.3) et l'attitude des acteurs clés par rapport aux évaluations indépendantes (3.1). Les différences se situent au niveau de la pertinence, de l'efficacité et de l'efficience en ce qui concerne le premier item, et uniquement au niveau de l'efficacité et de l'efficience en ce qui concerne le deuxième item.
En poussant plus loin l'analyse de ces deux composants, on constate en premier lieu que les interventions avec une TdC complexe ont des scores significativement plus élevés. Si nous faisons la comparaison entre les items sous-jacents, nous voyons que cette tendance se maintient, excepté pour l'item qui traite de l'identification des utilisateurs d'une évaluation et de leurs attentes (3.1.1.). Pour les interventions avec une TdC complexe, l'identification des utilisateurs d'une évaluation et de leurs attentes peut être plus difficile en raison, par exemple, du grand nombre d'acteurs différents avec des attentes distinctes dans les interventions de ce type, du fait qu'il y a, bien souvent, encore moins d'expérience avec une théorie de changement (plus) complexe si bien qu'il peut y avoir plus de facteurs incertains et que l'identification des utilisateurs et de leurs attentes est plus ardue, …

Le fait que les interventions avec une TdC moins complexe présentent, pour les critères CAD mentionnés, des scores significativement moins élevés au niveau du système S&E proposé (1.3) et dans ce cadre, au niveau de l'item concernant la traduction de la logique d'intervention sous-jacente dans le système S&E proposé, semble difficile à expliquer… Peut-être le fait que de telles interventions sont principalement mises en œuvre par des organisations moins expérimentées, si bien que des composants ou items techniques obtiennent aussi des scores plus faibles, joue-t-il un rôle ici[60].

Le score plus élevé pour les interventions avec une TdC plus complexe en ce qui concerne la compatibilité avec les attentes (3.1.3) peut, comme nous l'avons déjà suggéré dans la section 3.1 du chapitre précédent, trouver une explication dans le raisonnement suivant : il y a peu ou moins de problèmes du fait, comme le score pour 3.1.1 l'indique, que la définition des utilisateurs les plus importants et de leurs attentes ne se fait pas systématiquement. De cette manière, des différences éventuelles dans les attentes et les intérêts de ces utilisateurs ne sont pas mises en lumière.

Les scores plus élevés pour les interventions avec une TdC complexe en ce qui concerne l'item sur le fait de savoir si les acteurs clés sont demandeurs d'une – ou au moins intéressés par une – évaluation (3.1.2) peuvent avoir différentes causes (comme nous l'avons déjà évoqué à la section 3.1 du chapitre précédent) : les interventions avec des scores inférieurs pour cet item n'ont pas une attitude explicitement négative par rapport aux évaluations externes, un score plus faible pouvant suggérer qu'elles estiment que les évaluations externes n'apporteront aucune plus-value par rapport aux évaluations

[59] Les scores moyens figurent dans le tableau A2 7 de l'Annexe 7. Une courte description des tests statistiques correspondants a été reprise dans l'Annexe 7.

[60] L'échantillon est cependant trop limité pour pouvoir formuler des conclusions fiables à cet égard.

internes existantes et aux résultats du suivi déjà disponibles. Si par exemple les interventions avec une TdC moins complexe peuvent suivre plus facilement les résultats du monitoriong, il est possible que dans ces interventions, la plus-value d'évaluations indépendantes soit jugée plus faible.

Dans la plupart des interventions, les relations entre les acteurs clés sont saines. Il se peut que des interventions avec une TdC moins complexe présentent des scores inférieurs pour cet item (3.1.7) en raison du fait – comme nous l'avons évoqué précédemment – que les interventions avec une TdC complexe ont opéré une définition moins claire des utilisateurs et de leurs intérêts et attentes par rapport aux objectifs et au contenu de l'évaluation. Comme nous l'avons indiqué au point 3.1 du chapitre précédent, il est possible qu'une définition plus 'délibérée' des intérêts et attentes par rapport aux objectifs et au contenu de l'évaluation puisse donner une image plus nette des relations. D'un autre côté, si nous combinons le résultat pour '*les relations entre les acteurs clés sont saines*' et le score élevé (le plus élevé dans ce composant) pour l'item '*il est possible de contacter les acteurs clés sans que le risque existe qu'ils s'influencent mutuellement*', ceci pourrait également indiquer que dans les interventions avec une TdC plus complexe, vu l'incertitude, il y a un plus grand intérêt à l'égard d'opinions diverses (pour pouvoir en tirer plus d'enseignements).

Au niveau de l'évaluabilité de l'impact et de la durabilité, il n'y a pas de différences significatives entre les interventions avec une TdC complexe et avec une TdC moins complexe. Ce qui avait été décrit au sujet de ces niveaux dans les analyses précédentes, à savoir un développement moindre de ces niveaux comparativement aux trois autres critères CAD, peut être un élément d'explication. En outre, nous voudrions faire observer que, pour ce qui concerne les composants qui englobent l'analyse sous-jacente (1.1) et la logique d'intervention et la théorie de changement sous-jacentes (1.2), pour aucun des critères CAD pertinents nous n'avons trouvé de différences significatives entre les interventions avec une TdC complexe et avec une TdC moins complexe. Bien que les interventions avec une TdC complexe soient plutôt associées à plus de difficultés pour ce qui est du développement de la logique d'intervention et de la TdC, par exemple, il apparaît que dans la pratique ceci ne conduit pas à des analyses et/ou une logique d'intervention qui influencent négativement ces interventions. Pour toutes les interventions, les scores sont relativement bons pour ces composants. Il ressort également des visites de terrain que dans les interventions avec une TdC plus complexe, malgré les difficultés, on dépense beaucoup d'énergie pour élaborer les analyses et les théories de changement sous-jacentes et les traduire concrètement en une logique d'intervention consistante.

## 4.3 Comparaison de l'évaluabilité au niveau acteurs

Le tableau ci-dessous présente les scores au niveau *composants* et pour ce qui concerne les trois grandes dimensions du cadre d'étude. Comme pour l'analyse au niveau pays et au niveau complexité, il est important de signaler que des scores similaires au niveau composants peuvent dissimuler des différences au niveau *items*. Ces différences potentielles sont analysées en comparant les scores moyens par critère CAD au niveau composants et au niveau items entre les trois types d'acteurs.

**Tableau 14: Aperçu des scores d'évaluabilité selon le type d'acteur**

| | **CTB** | **ONG et syndicats** | **Autres** | Index d'éva-luabilité (°) |
|---|---|---|---|---|
| **Dimension 1 (plan d'intervention)** | **3,52** | **3,20** | **2,59** | **3,16** |
| 1.1 L'analyse sous-jacente | *4,10* | *3,75* | *2,90* | 3,65 |
| 1.2 La logique d'intervention et la théorie de changement | 3,00 | 2,98 | 2,90 | 2,96 |
| 1.3 Le système S&E proposé | 2,82 | 2,70 | 1,82 | 2,51 |
| 1.4 Consistance et adaptation de la logique d'intervention et la théorie de | 4,17 | 3,38 | 2,75 | 3,54 |

| | CTB | ONG et syndicats | Autres | Index d'éva-luabilité (°) |
|---|---|---|---|---|
| changement | | | | |
| **Dimension 2 (la pratique de mise en oeuvre)** | **2,97** | **3,09** | **2,70** | **2,96** |
| 2.1 Information de base quant à la mise en oeuvre de l'intervention | 2,86 | 3,03 | 2,62 | 2,88 |
| 2.2 Le système S&E dans la pratique | 2,95 | 3,10 | 2,78 | 3,04 |
| **Dimension 3 (contexte)** | **4,07** | **3,93** | **3,72** | **3,91** |
| 3.1 L'attitude des acteurs clés | 3,74 | 3,72 | 3,44 | 3,66 |
| 3.2 Le contexte plus large | 4,40 | 4,14 | 4,00 | 4,17 |
| **Score global évaluabilité (°°)** | ***3,41*** | ***3,30*** | ***2,86*** | ***3,23*** |

(°) Le score portant sur l'index d'évaluabilité a été calculé sur base d'indexes (*pas* sur base de moyennes). D'autre part, les scores des interventions avec TdC moins complexe et complexe sont des moyennes (*pas* des indexes). Ainsi il est possible que le score global d'évaluabilité diffère légèrement de la moyenne pondérée des scores des interventions 'CTB', 'ONG et syndicats' et 'autres' qui représentent respectivement 25%, 50% et 25% de l'échantillon. Ces différences sont cependant minimales et ne changent rien au contenu de l'analyse.
(°°) Pour le score global de l'évaluabilité, une moyenne pondérée a été calculée avec un poids identique pour les dimensions 1 et 2, tandis que le poids de la dimension 3 constitue seulement la moitié de celui des deux autres dimensions.

Comme le montre le tableau ci-dessus, il y a entre les différents acteurs ou 'canaux', comparativement aux deux paramètres précédents, des différences relativement plus grandes sur le plan de l'évaluabilité. Ces différences, bien entendu, ne se manifestent pas tant au niveau de la troisième dimension (influence du contexte), mais bien dans les deux autres dimensions et en particulier la dimension 1 (le plan d'intervention). Il ressort des données du tableau que l'évaluabilité des interventions CTB et ONG/syndicats est sensiblement plus élevée que pour les autres acteurs, ce qui ne veut pas dire que ces différences sont nécessairement présentes pour les interventions individuelles à l'intérieur de chaque groupe d'acteurs ; ceci vaut a fortiori pour le groupe 'Autres acteurs' qui englobe un large éventail d'organisations et de pratiques.

Une analyse plus précise du tableau ci-dessus et une comparaison des scores moyens qui ont été obtenus par critère CAD au niveau composants et items[61] nous permettent de mieux cerner les différences entre les acteurs. Les résultats des tests statistiques (voir l'annexe 7) montrent qu'entre les trois acteurs, il n'y a des différences significatives qu'au niveau du plan de l'intervention (partie 1). D'un autre côté, le contexte et la pratique de mise en œuvre de l'intervention pour ces trois canaux d'intervention ne présentent pas de différence significative quant à l'influence sur l'évaluabilité des interventions.

Il ressort en outre de l'analyse statistique que les différences concernant le plan de l'intervention se situent au niveau de l'analyse sous-jacente (1.1) et du système S&E proposé (1.3). La différence dans l'analyse sous-jacente (1.1) se manifeste principalement par des différences au niveau des items qui portent sur la délimitation et la description des groupes cibles (1.1.1), la description du rôle des groupes cibles et du rôle des acteurs (1.1.3 et 1.1.4) et la présence ou non d'une bonne analyse de genre (1.1.5). En ce qui concerne les trois premiers items mentionnés, les éléments les plus saillants sont les scores élevés des interventions CTB et les scores bas des interventions dans le canal d'intervention 'Autres'. Une explication réside peut-être dans le fait qu'à la CTB, il est demandé à toutes les interventions d'utiliser les mêmes canevas détaillés et bien élaborés pour établir les propositions d'intervention et le dossier technique et financier global. Dans ces canevas, une assez grande attention est aussi accordée à la description du groupe cible et du rôle que ces groupes cibles et/ou les différents acteurs assumeront dans l'intervention. En outre, tant à la CTB (via le processus MoreResults) que dans les ONG (via les initiatives des fédérations et les processus de changement dans bon nombre d'ONG), une grande attention est consacrée à l'amélioration de la gestion de projet. Ceci contraste avec le canal d'intervention "Autres" où, dans certains

[61] Les scores moyens figurent dans le tableau A2 4 de l'Annexe 7. Les résultats et une courte description des tests statistiques correspondants ont été repris dans cette même annexe.

cas, on n'exige même aucune proposition d'intervention et/ou analyse sous-jacente (pour la demande de fonds) et où les canevas utilisés apparaissent souvent moins complets et plus disparates. À cela s'ajoute le fait que pour certains de ces acteurs "Autres" (p. ex. les villes et communes), la coopération au développement ne constitue qu'une activité secondaire, dans laquelle ils ne sont pas vraiment spécialisés.

En ce qui concerne l'analyse de genre (1.1.5), on retrouve un score plus élevé pour le canal d'intervention 'ONG et syndicats'. Une explication possible est que ces acteurs sont, historiquement, plus enclins à atteindre les groupes cibles plus faibles socialement, où les relations de genre peuvent beaucoup plus facilement être associées à des mécanismes explicatifs sous-jacents, si bien que la nécessité d'une intégration de l'analyse de genre et de sa transposition dans le système S&E y est plus évidente. Le score relativement faible des interventions CTB peut trouver son origine dans le fait que, lorsque les interventions étudiées ont été planifiées et mises en œuvre, le genre ne faisait l'objet que d'une attention sporadique. L'attention pour le genre dans le manuel MoreResults n'est pas fondamentalement différente par rapport aux directives antérieures.

Entre les trois canaux d'intervention, aucune différence significative n'a en outre été trouvée dans les scores pour les items concernant le bien fondé de l'intervention (1.1.2), le lien entre l'analyse sous-jacente et les objectifs de l'intervention (1.1.6) et la position de l'intervention par rapport à la politique sectorielle locale (1.1.7). C'est une indication supplémentaire permettant d'affirmer que de bons canevas jouent un grand rôle dans l'élaboration du plan, mais que derrière les interventions des trois canaux d'intervention distincts se cache bel et bien un raisonnement dans lequel des objectifs d'intervention logiques sont préétablis et dans lequel il est tenu compte du contexte de politique dans lequel s'inscrit l'intervention. Par ailleurs, aucune différence significative n'est observée entre les trois canaux d'intervention en ce qui concerne la logique d'intervention et la théorie de changement (1.2). Cette constatation indique elle aussi qu'il n'y a, par canal d'intervention, aucune différence au niveau de la compétence technique pour ce qui est de cadrer les interventions dans une théorie de changement et d'élaborer une logique d'intervention.

En ce qui concerne les différences identifiées pour les trois canaux d'intervention au niveau du système S&E proposé (1.3), nous observons trois "tendances" distinctes. Premièrement, les interventions du canal d'intervention "Autres" présentent un score significativement inférieur au niveau du composant 'système S&E proposé', que l'on retrouve également dans les items concernant l'opérationnalisation des résultats de l'intervention (1.3.1), la description des ressources personnelles et financières du système S&E (1.3.7), la présence ou non ou l'utilisation d'un MIS (1.3.8) et le système S&E en tant qu'opérationnalisation consistante de la logique d'intervention (1.3.3). Comme indiqué précédemment, d'autres conditions et exigences s'appliquent pour ce type de canaux d'intervention en ce qui concerne les rapports et leur format sur le plan du contenu. Dans ce type de canaux d'intervention, il n'est pas toujours exigé que des objectifs opérationnalisés soient préétablis, de même qu'il n'est pas demandé qu'ils soient suivis de façon systématique. Il n'est pas étonnant non plus que des aspects pratiques comme les ressources disponibles pour le système S&E soient moins mises en lumière, étant donné que dans les canaux d'intervention de ce type, il est beaucoup moins question, dans les interventions, d'une culture et d'une pratique S&E existantes. Ce constat général n'exclut pas, bien évidemment, que certaines interventions individuelles aient proposé et élaboré de bonnes pratiques en matière de S&E. On remarquera à cet égard que les scores relatifs au système S&E dans la pratique sont beaucoup plus proches qu'en ce qui concerne le système S&E proposé.
Deuxièmement, nous constatons que les interventions dans le canal d'intervention "ONG et syndicats" obtiennent des scores significativement plus élevés en ce qui concerne la description claire de l'approche de suivi des hypothèses (1.3.5). Une explication possible est que pour les interventions des ONG et des syndicats, la politique des pouvoirs publics fait partie du contexte externe et que dans ces interventions, on est conscient que la politique des pouvoirs publics fait partie des hypothèses externes qui peuvent influencer les interventions. D'un autre côté, pour les interventions bilatérales, la politique des

pouvoirs publics fait beaucoup moins partie des hypothèses externes (mais plutôt des risques internes) et les autres hypothèses externes qui peuvent potentiellement influencer les interventions sont beaucoup plus difficiles à identifier. Les interventions dans le canal d'intervention "Autres" sont, quant à elles, moins liées à une politique des pouvoirs publics et peuvent éprouver des difficultés à identifier d'autres hypothèses externes.

Troisièmement, les interventions dans le canal d'intervention "CTB" affichent des scores plus élevés en ce qui concerne la description claire de la manière dont le système S&E s'articule par rapport aux systèmes S&E locaux/nationaux (1.3.9). Cette différence s'explique aisément si l'on considère les modalités de mise en œuvre selon lesquelles les interventions CTB sont exécutées et où les systèmes de suivi correspondants constituent un point d'attention plus explicite.

# 5 Conclusions et recommandations

## 5.1 Principales conclusions

Les conclusions et enseignements qui suivent sont présentés en deux parties. La première partie est essentiellement un résumé des principales constatations, tandis que la deuxième est de nature plus analytique.

### 5.1.1 Synthèse des principaux résultats et constatations

#### Signification et importance de l'« évaluabilité » dans la pratique du développement

L'évaluabilité est définie comme la mesure selon laquelle une activité ou un programme peut être évalué de façon fiable et crédible. Dans ce cadre, la détermination de la "faisabilité" d'une évaluation est étendue à l'examen de l'*opportunité* d'une évaluation. Ceci implique que l'évaluabilité se concentre avant tout sur trois dimensions : le (la qualité du) plan d'intervention, (la qualité de) la mise en œuvre de l'intervention (système S&E inclus) et le rôle de l'environnement plus large de l'intervention.

Le premier intérêt – et le plus évident – de l'évaluabilité part du constat selon lequel la mise en œuvre d'un test d'évaluabilité (ou d'une appréciation de l'évaluabilité) ne requiert qu'une fraction des moyens qui sont nécessaires pour la mise en œuvre de l'évaluation proprement dite, alors que la plus-value potentielle d'un tel test est appréciable. Autrement dit, une appréciation de l'évaluabilité permet d'établir si une évaluation (à un moment déterminé, dans un contexte déterminé) est souhaitable et réalisable et si oui, sous quelles conditions.

Néanmoins, une utilisation adéquate de l'« évaluabilité » et la mise en œuvre d'une appréciation de l'évaluabilité ne doivent pas nécessairement être associées à l'organisation des évaluations mais ont une *utilité plus large* qui peut être bénéfique à la gestion des interventions de développement dans son ensemble. Une analyse de l'évaluabilité peut ainsi améliorer le plan d'intervention, apporter une contribution substantielle à la conception et l'élaboration d'un système S&E, fournir des inputs intéressants en ce qui concerne l'opportunité, le timing, l'approche et les objectifs d'une évaluation et contribuer de cette manière à l'utilité pratique et à l'utilisation effective des résultats de l'évaluation, et donc à l'utilité finale de l'évaluation.

Ce qui précède révèle l'intérêt à la fois du concept d'évaluabilité et de l'instrument "test d'évaluabilité" : leur application potentielle va au-delà (de l'organisation) des évaluations et concerne la gestion des interventions dans son ensemble. Toutefois, l'étude a permis de constater que le concept *évaluabilité* et l'instrument *test d'évaluabilité* étaient, à ce jour, peu connus et utilisés dans la coopération belge au développement. Certains éléments d'un test d'évaluabilité sont mis en pratique ici et là (sans qu'ils soient nommés comme tels), mais nulle part il n'est question d'une application systématique.
Cette constatation est significative, car l'« évaluation » a peu à peu gagné en importance ces dernières décennies, jusqu'à faire partie intégrante de la pratique de gestion au niveau des interventions, des programmes et de la coopération au développement dans son ensemble : la pertinence et l'utilité des évaluations ne sont pas contestées et les acteurs ne peuvent tout simplement plus se permettre de ne *pas* évaluer. Autrement dit,

évaluer est devenu une obligation. Bien que cette évolution soit globalement positive, elle porte aussi en elle le risque que les évaluations soient ramenées au rang d'exercices rituels, sans implication authentique des acteurs clés. Une utilisation délibérée de la notion d'« évaluabilité » et de l'instrument « test d'évaluabilité », avec en corollaire la possibilité d'émettre des jugements fondés quant à l'opportunité d'une évaluation, peut constituer, dans ce contexte, un important instrument pour améliorer – et rendre plus pertinente et réaliste – la manière d'exercer le rôle et la fonction des évaluations dans la coopération au développement.

## L'évaluabilité envisagée dans sa globalité

Les 40 interventions analysées obtiennent un score global (sur la base des 62 items du cadre d'étude) qui se situe légèrement au-dessus du score moyen (voir tableau 3). Même si nous ne pouvons pas donner une signification absolue aux scores, le score global est une bonne indication de la principale constatation de ces études, à savoir que **les interventions ont dans l'ensemble un certain nombre de points forts, mais aussi un grand nombre de points à travailler si elles veulent améliorer leur évaluabilité.** Il ressort également que la distribution (statistique) des scores est proche d'une distribution gaussienne (normale), mais avec une dispersion importante : les interventions avec les scores les plus bas n'atteignent pas – sur le plan du score – la moitié de celles avec les scores les plus élevés. Étant donné que l'« évaluabilité » est fortement liée à la pratique de gestion de l'intervention, cette constatation implique qu'il existe encore, sur le plan de la gestion, de grandes différences parmi les acteurs belges et les types d'interventions, malgré le pilotage de l'autorité qui assure le financement.

Sur les trois dimensions étudiées (plan d'intervention, pratique de mise en œuvre, contexte), la dernière affiche des scores sensiblement plus élevés que les autres. Ceci est une illustration supplémentaire de la marge d'amélioration de ces dimensions, sur lesquelles les acteurs ont le plus de prise (plan d'intervention et pratique de mise en œuvre). Ensuite, le score semble indiquer que – tout au moins pour les 4 pays étudiés – les défis au niveau du contexte n'hypothèquent par lourdement l'évaluabilité. À l'analyse, il apparaît cependant que cette interprétation doit être nuancée, car les limitations éventuelles au niveau du contexte n'ont pas pu être analysées en profondeur.

Si nous observons les scores globaux par critère d'évaluation, nous remarquons d'emblée les scores plus faibles pour la durabilité et surtout pour l'impact : malgré l'attention croissante accordée à la durabilité, celle-ci reste insuffisamment intégrée dans les systèmes de gestion, tandis qu'au niveau de l'impact, une conjonction de facteurs entraîne une évaluabilité difficile. De l'autre côté, l'efficacité et surtout l'efficience présentent des scores nettement supérieurs à la valeur moyenne de l'échelle développée, ce qui est – pour l'ensemble des dimensions et composants – une indication de qualité dans la gestion de l'intervention, notamment en ce qui concerne le suivi et l'évaluation.

## Le plan d'intervention

Le bon score global pour **l'analyse sous-jacente** révèle une certaine tradition et aptitude pour ce qui est de l'élaboration d'analyses, ceci résultant à la fois de processus internes et d'exigences des bailleurs de fonds pour qui de telles analyses forment une composante importante dans l'appréciation des demandes de subsides. Dans bien des cas, les initiateurs peuvent aussi s'appuyer sur les expériences du passé (interventions subséquentes). Un constat positif important est le fait que dans de nombreuses analyses, une grande attention est portée au contexte de politique, ceci résultant en partie d'une amélioration du réseautage et des activités de plateforme parmi les acteurs belges.
D'un autre côte, il y a aussi sur le plan de l'analyse quelques points problématiques notables : le peu d'attention, au niveau des groupes cibles, pour le genre et la (les facteurs de) différenciation sociale en général, l'attention limitée pour les acteurs clés (autres que les groupes cibles), ou encore le resserrement des analyses qui, souvent,

servent avant tout à justifier les choix stratégiques et les choix de politique plutôt que d'en former la base. Ces lacunes influencent négativement l'évaluabilité en ce sens qu'elles compliquent l'appréciation de la pertinence des choix qui sont faits en matière de politique et qu'il est plus difficile d'évaluer qui est effectivement atteint (en comparaison avec le planning) ou si des effets d'éviction ont joué au niveau du groupe cible.

Le composant **logique d'intervention et théorie de changement** enregistre des scores relativement faibles, ce qui est dû à l'attention réduite accordée aux niveaux supérieurs dans la chaîne moyens-fin. Dans la pratique, l'attention se porte avant tout sur le niveau mise en œuvre (lors du screening, c'est surtout ce niveau qui est pris en considération) et il semble y avoir peu d'incitants, et donc peu d'intérêt à dépasser ce niveau : non seulement le niveau mise en œuvre a une utilité plus grande et plus directe pour les organisations concernées, mais l'autorité assurant le financement est avant tout intéressée, de facto, par la justification (en ce qui concerne la bonne utilisation des moyens), et les canevas développés accordent par exemple peu d'attention au niveau impact. Une difficulté supplémentaire est que les impacts, pour autant qu'ils soient formulés, se situent à une très grande distance dans la chaîne moyens-fin, si bien qu'il se crée un '*missing middle*' dans la théorie de changement. En raison des défis intrinsèques existants, du manque d'incitants externes et d'une connaissance et une compréhension limitées des méthodes et opportunités existantes (et réalistes) de l'évaluation de l'impact, il existe, déjà dans la phase de préparation, d'importantes difficultés qui hypothèquent l'évaluabilité de l'impact et la mise en œuvre effective d'évaluations d'impact (c.-à-d. en ce qui concerne les outcomes et leurs effets directs) à un stade ultérieur.
D'un autre côté, la focalisation sur la 'mise en œuvre' assure un bon score pour le critère d'efficience, ce qui se répercute aussi effectivement dans la pratique (p. ex. dans la bonne liaison entre l'utilisation des moyens et les outputs).

La qualité de **la proposition S&E dans la proposition d'intervention** obtient des scores faibles. Les causes de ce résultat sont diverses : le manque d'incitants pour investir résolument, ex ante, dans la description du système S&E ; le fait que l'attention pour le S&E n'a fortement augmenté que récemment (de nombreux acteurs ayant fourni des efforts substantiels pour élaborer une politique S&E) mais ne s'est pas encore répercutée au niveau intervention ; ou encore le choix de certains acteurs de n'investir dans l'élaboration d'un système S&E qu'au début d'une intervention. Du point de vue de l'évaluabilité, la disponibilité – ex ante – d'un plan S&E n'est importante que dans la mesure où cela implique, a posteriori, une bonne pratique S&E. Dans la pratique, en effet, une bonne attention préalable est souvent annonciatrice d'une bonne pratique, mais il y a aussi des interventions où un score faible ex ante est corrigé après coup par une bonne pratique. Par ailleurs, des propositions S&E bien élaborées (sur papier) s'avèrent aussi importantes à la lumière de la forte rotation de personnel et des nombreux changements de tâches dans bon nombre d'interventions. Enfin, il est clair que dans bien des cas, les points problématiques (p. ex. une faible attention pour l'impact et le genre ou l'absence de description spécifique et d'attribution de moyens pour le suivi et surtout pour l'évaluation) reviennent par la suite. Ceci vaut – a fortiori – pour le constat selon lequel il n'est que rarement question d'un véritable *système* S&E : dans de nombreux cas, des composantes d'un tel système sont proposées, mais il reste beaucoup à faire pour garantir une bonne cohérence interne.

Les **changements dans la logique d'intervention et la théorie de changement pendant la mise en œuvre** sont, en général, bien signalés et argumentés, ce qui est une constatation positive du point de vue de l'évaluabilité. Bien souvent, toutefois, ces changements ne sont pas répercutés de manière consistante dans le système S&E, ce qui révèle que l'ancrage institutionnel de ce système reste faible.

## La pratique de mise en œuvre

L'**information de base relative au progrès dans la mise en œuvre de l'intervention** est relativement disponible mais – notamment en raison de la langue ou, parfois, de la complexité – n'est pas toujours accessible pour (et connue de) tous les

acteurs clés. Par ailleurs, cette information reflète l'information qui est rendue disponible par le plan d'intervention : elle se concentre surtout sur le niveau mise en œuvre, reste superficielle pour ce qui est des aspects de différenciation sociale (y compris le genre) et accorde peu d'attention aux niveaux supérieurs dans la chaîne moyens-fin. En conséquence, l'information laisse présager une tendance "*inward looking*" dans la gestion de l'intervention, avec peu d'attention pour les hypothèses externes, la possibilité de développer un counterfactual et l'approche et la qualité du processus de collecte de données. Le score assez faible pour cet item doit cependant être nuancé, car les visites de terrain ont révélé, dans un grand nombre de cas, des bonnes pratiques dont on ne trouvait aucune trace dans les documents de base.

Le **système S&E dans la pratique** est sans doute le composant qui influence le plus l'évaluabilité d'une intervention. Des faiblesses à d'autres niveaux peuvent en effet, jusqu'à un certain point, être compensées par un système S&E cohérent qui fonctionne bien ; à l'inverse, si le système S&E fonctionne mal, cela compliquera l'évaluabilité pratique d'une intervention. À la lumière de cette observation, le fait que la pratique S&E *effective* obtient des scores sensiblement supérieurs au système S&E *proposé* est une constatation positive.
À l'analyse, il apparaît que la pratique S&E est dans bien des cas un 'travail en chantier' qui se limite encore trop au niveau opérationnel pour des raisons qui ont déjà été partiellement évoquées ci-dessus et qu'il faut attribuer à des manquements dans le plan d'intervention et à l'attention suffisante mais somme toute très récente portée au S&E, ce qui fait que les décisions en matière de politique sur ce plan n'ont pas encore été répercutées au niveau local. Il semble aussi qu'interviennent une série de mécanismes qui influencent les caractéristiques de la pratique S&E : la relation étroite entre le suivi opérationnel et les tâches essentielles du personnel du programme (contrairement au suivi axé sur le résultat qui est souvent plus exigeant et est "plus éloigné") ; l'absence d'indicateurs au niveau impact et (parfois) au niveau outcome, si bien que ces niveaux sont insuffisamment mis en lumière dans la pratique S&E ; l'absence de développement des compétences spécifiques en matière de S&E (ou le manque d'attention à cet égard) ; le fait que le S&E constitue pour beaucoup une responsabilité supplémentaire ; et le manque de temps, de moyens et de procédures pour utiliser correctement les résultats S&E pour l'analyse approfondie, la prise de décision et le réajustement. Ces points faibles sont toutefois compensés en partie par une culture d'organisation centrée sur la réflexion et l'apprentissage.
D'autre part, il apparaît que la politique et la pratique S&E sont jusqu'ici centrées principalement sur le "suivi" au détriment de l'évaluation. Il y a à cela plusieurs explications : le fait que la politique S&E doit encore prendre forme dans la pratique, si bien qu'il est logique que l'attention se porte initialement sur le suivi ; les exigences méthodologiques plus importantes lors de l'évaluation comparativement au suivi ; le doute (notamment dans les organisations avec un système de suivi performant et dans les interventions complexes) quant à la plus-value potentielle d'évaluations externes, si bien que l'on prévoira peu de moyens pour de telles évaluations (voir également ci-après).
Une dernière constatation importante est liée à l'attitude résolument '*inward looking*' qui caractérise la pratique S&E de nombreux acteurs belges du développement et qui se traduit surtout par un rôle limité dans la politique et la mise en œuvre S&E pour les acteurs de l'intervention (autres que l'organisation responsable et l'équipe de l'intervention), une redevabilité qui se réduit à la redevabilité vis-à-vis du bailleur de fonds et une attention limitée pour l'adéquation du S&E avec les systèmes locaux et nationaux (y compris celui du partenaire) et pour le rôle éventuel que les acteurs belges peuvent/doivent jouer dans la mise en place et l'amélioration de ces systèmes.

En résumé, la pratique de mise en œuvre actuelle comporte à la fois des possibilités et des limitations par rapport à l'évaluabilité ; dans ce cadre, il semble important de garder à l'esprit que cette étude fournit un instantané d'une pratique sujette à des changements importants.

## L'influence du contexte de l'évaluation

L'**attitude des acteurs clés** par rapport aux évaluations (indépendantes) est un important facteur d'évaluabilité. L'étude a révélé que cette attitude ne constituait un obstacle direct que dans de rares cas. En règle générale, les acteurs clés se montrent positifs et intéressés à l'égard des évaluations externes et connaissent le rôle et l'attitude (p. ex. respect pour l'autonomie de l'évaluateur, pas de tentatives de l'influencer) que l'on attend d'eux. Il s'avère toutefois, ici aussi, que la réalité est un peu plus complexe qu'il peut sembler à première vue : dans la pratique, en effet, il apparaît que seul un nombre très limité d'acteurs est impliqué dans le processus d'évaluation et que les intérêts et les attentes d'un large éventail d'acteurs ne sont pas pris en compte dans les évaluations (en raison notamment d'une politique S&E imprécise, d'une pratique d'évaluation trop faible et d'un scepticisme quant à la plus-value des évaluations externes). Lorsque les intérêts et les attentes de ces acteurs ne sont pas pris en compte, les choix fondamentaux en ce qui concerne l'évaluation sont généralement déterminés par les '*usual suspects*' qui, de cette manière, peuvent orienter l'évaluation avant qu'elle ait commencé dans la pratique.

Les résultats concernant l'influence du **contexte plus large** donnent une image similaire. À première vue, les évaluations indépendantes semblent pouvoir être exécutées sans difficultés notables et les défis qui se présentent semblent pouvoir être résolus de manière adéquate. Il est possible, toutefois, que les facteurs contextuels qui mettent (peuvent mettre) l'évaluabilité sous pression aient aussi influencé la présente étude (par exemple en fournissant des réponses socialement souhaitables), d'autant que cette étude (comme bon nombre d'évaluations) a été réalisée dans des délais très serrés. En d'autres termes, il est possible que l'absence de problèmes majeurs puisse s'expliquer par une focalisation sur les dimensions 'techniques' de l'évaluabilité au détriment des dimensions politiques qui, pour diverses raisons, n'ont pas été intégrées de manière explicite.

## L'influence du contexte spécifique aux pays sur l'évaluabilité

Les scores d'évaluabilité par pays présentent très peu de différences entre eux, ce qui implique que d'autres paramètres sont sans doute plus importants. On n'observe – pour des raisons évidentes – des différences relativement grandes qu'en ce qui concerne le contexte ; néanmoins, l'influence du contexte sur l'évaluabilité 'technique' n'est pas si grande, dans le sens où elle n'engendre pas d'obstacles majeurs. Du reste, un contexte institutionnel donné peut influer de manière tant positive que négative sur différents aspects de l'évaluabilité. Dans le cas du Rwanda, par exemple, la culture résolument axée sur la performance a une influence positive sur la demande de suivi ou l'élaboration effective de systèmes de suivi (centrés principalement sur la redevabilité ascendante), tandis que cette même focalisation sur la performance et les *quick-wins* peut entraver le signalement des échecs et des changements, ce qui exerce simultanément une influence négative sur la fonction d'apprentissage/rétroaction et met l'évaluabilité sous pression.

## L'influence du degré de complexité des interventions sur l'évaluabilité

Il n'y a pas de différences fondamentales entre l'évaluabilité des interventions avec une TdC 'complexe' et avec une TdC 'moins complexe'. Les interventions avec une TdC complexe affichent même des scores légèrement supérieurs, sans doute parce que les acteurs de ces interventions investissent plus dans l'analyse et l'élaboration de systèmes et de pratiques S&E, considérant – à tort ou à raison – que ces interventions sont plus difficiles à financer et que leurs résultats sont plus difficiles à démontrer. Par ailleurs, les petites différences entre ces deux types d'interventions peuvent aussi être liées au fait que certains paramètres de complexité (comme le nombre d'acteurs impliqués) n'ont pas été pris en compte lors de la composition des deux groupes.

### L'influence du canal de financement sur l'évaluabilité

Les scores d'évaluabilité par canal de financement (type d'acteur belge) présentent des différences plus marquées que pour la nature des interventions et le pays. On relève en particulier des différences importantes entre 'bilatéral/ONG/syndicats' d'une part et les 'autres acteurs' d'autre part (meilleurs scores pour le premier groupe), même s'il y a aussi des exemples de bonnes pratiques dans ce dernier groupe. La principale explication de ce constat réside probablement dans les exigences externes plus basses (pour les 'autres acteurs') par rapport au plan d'intervention et à la pratique de mise en œuvre, ceci étant renforcé par le fait que pour une partie des acteurs appartenant à ce groupe, la 'coopération au développement' ne constitue pas une tâche principale. D'un autre côté, la CTB et les ONG/syndicats ont aussi, en tant que 'secteur', beaucoup investi dans un S&E amélioré (principalement le suivi).

## 5.1.2 Analyse

Si nous comparons entre eux les scores relatifs aux cinq critères d'évaluation, il apparaît d'emblée que, pour presque tous les items (et donc aussi pour tous les composants et dimensions), le critère d'efficience affiche les scores les plus élevés, suivi par l'efficacité, la durabilité et l'impact. Le score pour la pertinence se situe généralement au milieu. Le fait que l'on retrouve presque à chaque fois le même schéma dans les scores est évidemment lié aux différents degrés de difficulté – dans toutes les phases du cycle d'intervention – de l'évaluation de ces critères. Il est difficile, sur ce plan, de généraliser, mais on peut affirmer que la durabilité et plus encore l'impact sont plus difficiles à évaluer que les trois autres critères. L'évaluation de l'impact pose des exigences méthodologiques élevées, tandis que pour l'évaluation de la durabilité, la difficulté est le plus souvent liée au défi que constitue la formulation de jugements fondés par rapport à une situation qui ne se produira que dans le futur.

Mais il y a d'autres raisons encore qui peuvent expliquer la différence entre les scores des cinq critères. Premièrement, les effets de la direction donnée par les bailleurs de fonds se font surtout sentir au niveau de l'efficience. Les bailleurs (et en particulier la DGD) se préoccupent avant tout de l'utilisation correcte des fonds publics mis à disposition et ont imposé à cet effet d'importantes conditions pour la gestion des interventions qu'ils financent. Les procédures et les canevas décrivent dans une large mesure les priorités du bailleur de fonds, même si les acteurs concernés peuvent aussi, à travers leur dialogue avec les pouvoirs publics, y inclure certains accents. De manière analogue, les canevas utilisés (notamment pour les propositions d'intervention et les rapports d'avancement) semblent jouer un rôle important dans la détermination de l'évaluabilité du fait, par exemple, qu'ils mettent fortement l'accent sur les aspects liés à l'efficience, mais accordent peu d'attention à l'impact. Du reste, ces canevas sont aussi inspirés jusqu'à un certain point par le souhait de donner corps à une gestion axée sur les résultats, par exemple en exigeant que les dépenses prévues au budget soient mises en relation avec les outputs. Par ailleurs, le bon score pour l'évaluabilité de l'efficience s'explique aussi, certainement, par les efforts importants entrepris notamment par la CTB et les ONG pour développer et mettre en œuvre leurs systèmes S&E. Ces efforts sont la conséquence de processus qui ont cours depuis déjà tout un temps au sein de ces organisations, mais qui constituent aussi une réaction au screening programmé des ANG, qui aura lieu en 2016 et qui examinera entre autres la qualité des systèmes S&E[62]. En ce qui concerne le développement de systèmes S&E, il s'agit dans bien des cas d'initiatives assez complexes qui évoluent de manière graduelle, si bien que les changements au niveau intervention ne sont introduits que progressivement. Dans ces processus, assez logiquement, une approche 'bottom-up' est suivie, le niveau mise en œuvre (inputs –

[62] Outre ce screening, une *certification* des systèmes S&E des acteurs de la coopération belge est également prévue. Cette certification fait suite à la réponse donnée par la direction de la DGD à la méta-évaluation des programmes des ONG, exécutée elle aussi pour le compte du SES (Méta-évaluation des programmes des acteurs non gouvernementaux, juillet 2013). Le rapport contient une recommandation concernant l'organisation et la mise en œuvre de cette certification (voir chapitre 5.2).

activités – outputs) étant le premier qui entre en ligne de compte ; du reste, c'est aussi le niveau où l'utilité directe pour les organisations concernées est la plus tangible, et celui qui intéresse le plus la DGD.
Il n'est cependant pas évident, pour diverses raisons, que les systèmes S&E intégreront progressivement, de façon "automatique", les niveaux supérieurs de la chaîne moyens-fin et que la fonction d'évaluation, entre autres, sera développée aussi fortement que le suivi. Car les incitants et les conditions préalables positives qui existent actuellement pour le S&E au niveau mise en œuvre (utilité directement démontrable, focus mis traditionnellement sur l'aspect opérationnel, pression de la DGD, exigences relativement peu élevées pour l'organisation et la mise en œuvre) sont moins présents – voire pas du tout – pour le S&E en ce qui concerne les outcomes et les impacts[63]. Une conclusion importante qui en découle est qu'en dépit des progrès enregistrés dans le développement de systèmes S&E ces dernières années, à politique et contexte inchangés, l'amélioration de l'évaluabilité (concernant en premier lieu les autres critères que l'efficience) n'est absolument pas assurée.

Une autre constatation importante de cette étude est que les imperfections dans le plan d'intervention continuent généralement à se faire sentir pendant la mise en œuvre et de cette manière, influencent directement ou indirectement le niveau d'évaluabilité. Ceci vaut notamment pour :

- l'intégration du genre et autres facteurs de différenciation sociale : s'il n'y a, initialement (pendant l'analyse du contexte et des groupes cibles), aucune attention pour le genre, la probabilité est grande que ce sera également le cas pendant l'élaboration de la baseline et par la suite du système S&E, et, au cours de la mise en œuvre, le genre restera sous-exposé et ne sera pas repris comme point d'attention dans les systèmes de suivi et les évaluations ;
- une faible attention ex ante pour le niveau impact (par exemple par l'absence des niveaux supérieurs ou un '*missing middle*' dans la TdC) qui se répercute dans une logique d'intervention incomplète (comme décrit dans le cadre logique) et, partant, dans le système S&E proposé et par la suite dans la pratique de mise en œuvre S&E, avec de facto de nombreuses interventions qui n'accordent pas la moindre attention au niveau impact et qui n'intègrent pas l'analyse d'impact dans leurs évaluations. De cette manière, des interventions ou des programmes plus étendus peuvent se poursuivre pendant plusieurs cycles de financement sans qu'une réflexion ait lieu sur les effets potentiels de l'action et a fortiori, sans que ces effets soient mis en lumière ;
- le fait que les caractéristiques, intérêts, … des acteurs et autres groupes concernés sont, ex ante, insuffisamment mis en lumière, implique qu'ils ne sont pas – ou pas suffisamment – impliqués dans l'élaboration des systèmes S&E, pas – ou pas suffisamment – pris en compte dans la baseline ni, dès lors, dans le système S&E dans la pratique, où ils ne jouent aucun rôle. Plus loin dans le cycle d'intervention, cela implique aussi que la probabilité que les intérêts et attentes de ces acteurs soient pris en compte dans la préparation et la mise en œuvre des évaluations est assez limitée, avec des conséquences néfastes pour l'évaluabilité, mais aussi pour l'utilité finale des évaluations ;
- si la façon selon laquelle le système S&E de l'intervention s'articule par rapport au système S&E national/local n'est pas décrite dans la proposition, il n'est pas étonnant que le système S&E de l'intervention, dans la pratique, ne soit pas non plus accordé avec ces systèmes nationaux ou locaux.

Comme l'illustrent, entre autres, les exemples ci-dessus, l'étude a montré que la qualité du plan d'intervention – ou mieux encore, de la phase de conception – est un facteur qui influence fortement, directement et indirectement, l'évaluabilité. Une bonne phase de

[63] La 'Note stratégique Résultats de développement' élaborée récemment met toutefois fortement l'accent sur le niveau outcome et indique par exemple (p. 3) que les outputs ne sont *pas* considérés comme des résultats de développement. Si cette note stratégique est bien suivie et est traduite de façon conséquente en changements sur le plan des exigences et des procédures, elle pourrait entraîner un revirement à cet égard.

conception est en effet annonciatrice, bien souvent, d'une gestion d'intervention de bonne qualité. Il apparaît que des investissements dans une bonne préparation de l'intervention sont récupérés par la suite, tandis que corriger par après des faiblesses initiales s'avère plus difficile que prévu au départ ; autrement dit, le plan initial est la référence sur laquelle s'appuie la pratique. Quant à savoir comment concilier ce constat avec la volonté actuelle de tendre – à juste titre – vers une simplification administrative, nous y reviendrons dans nos recommandations.

Il semble par ailleurs se créer une différence graduelle entre les conséquences d'une phase de conception faible pour le suivi d'une part, pour l'évaluation de l'autre. Comme notre analyse l'a démontré, pour une série de raisons, la pratique S&E affiche, du point de vue de l'évaluabilité, des scores nettement plus élevés que le plan S&E. À l'évidence, il est donc possible à ce niveau – principalement opérationnel – d'apporter assez facilement des corrections, même si les lacunes de la phase initiales continuent à se faire sentir (voir supra). Ces lacunes s'avèrent toutefois plus conséquentes pour la fonction d'évaluation, parce qu'il est plus difficile d'apporter des corrections et que des lacunes dans le plan d'intervention et la mise en œuvre de l'intervention impliquent que des aspects importants d'une intervention (atteindre effectivement les groupes cibles initiaux, effets de l'intervention sur différents groupes sociaux, réalisation des hypothèses et des risques) ne peuvent (pratiquement) pas être évalués et parfois même, restent totalement en dehors du champ de vision des évaluations. Dans ce cas, la conjonction des facteurs mentionnés ci-dessus peut avoir pour effet que les évaluations 'indépendantes' soient, de facto, fortement dirigées (ou tout au moins déterminées) de l'intérieur, si bien que – volontairement ou non – les vides qui subsistent ne sont pas détectés et les sujets controversés sont écartés ou restent sous-exposés. En d'autres termes, il y a un risque que les évaluations soient uniquement axées sur la réalité telle qu'elle est définie ou interprétée par l'intervention en question. Le fait que l'utilité globale et la pertinence des évaluations (en tant que composante obligatoire d'une bonne pratique de développement) et l'opportunité des évaluations sont rarement remises en question, peut encore renforcer ce risque. Dans ce genre de situation, même des évaluateurs expérimentés auront du mal à remettre en question, s'il y a lieu, les orientations données à l'évaluation ; dans bien des cas, ils n'en auront même pas la possibilité ou n'y seront pas enclins.

En résumé, les nombreuses initiatives visant à améliorer la gestion des interventions et à contribuer ainsi à une meilleure évaluabilité et (vraisemblablement) à une performance accrue en termes de développement, semblent avoir engrangé des résultats notables. Il subsiste néanmoins d'importantes lacunes. Ainsi, l'absence presque totale de vision de l'impact et de la durabilité des effets des interventions financées en grande partie avec de l'argent public, pose problème d'un point de vue sociétal. Il en va de même du développement incomplet de la fonction d'évaluation, qui a pour effet que les évaluations indépendantes doivent se dérouler de facto dans un cadre restrictif et n'arrivent dès lors que partiellement à atteindre leurs objectifs sur le plan de l'apprentissage et de la redevabilité.

## 5.2 Recommandations

**Remarque préalable**

Dans le cahier des charges de l'étude, au point B3, il est indiqué que l'étude doit être utile pour trois groupes distincts : le SES, les services de la DG-D et les Attachés et les partenaires de la coopération bilatérale et non gouvernementale. Dans la vision de l'équipe d'étude, chacun de ces groupes a un intérêt et une responsabilité dans les efforts visant à une meilleure évaluabilité et l'assument, idéalement, au départ d'un cadre et de directives définis et endossés *en commun*, à l'intérieur desquels chaque groupe agit dans son propre rôle et avec sa spécificité. Ce faisant, l'évaluabilité et l'appréciation de l'évaluabilité forment un point d'attention supplémentaire dans la tradition de concertation qui existe déjà entre les différents acteurs. À ce titre, les

recommandations stratégiques sont importantes pour toutes les parties. Quant aux recommandations opérationnelles, elles sont surtout importantes pour les partenaires de la coopération.
L'équipe d'étude a conscience que certaines des recommandations formulées sont assez ardues, si pas à terme, du moins au début. Dès lors, elles ne peuvent être appliquées que si la charge liée à la gestion qui incombe aux acteurs concernés peut être allégée en conséquence sur certains points. Dans ce cadre, les tentatives visant à une simplification administrative sont une condition nécessaire, mais insuffisante. Il est important également de *concevoir différemment* l'élaboration des propositions d'intervention, des rapports de mise en œuvre, etc., en portant l'attention sur les résultats de développement (outcomes, impact) plutôt que sur les niveaux opérationnels (moyens, activités, outputs). Cette considération se reflète dans plusieurs des recommandations développées ci-après.

### 5.2.1 Recommandations stratégiques

1. *La présente étude a tenté de concrétiser la notion et la pratique de l'appréciation de l'évaluabilité dans le contexte de la coopération belge au développement. Ceci a révélé que le concept et surtout l'appréciation effective de l'évaluabilité étaient, à ce jour, largement méconnues dans la coopération. Ensuite, l'étude a démontré que l'appréciation de l'évaluabilité pouvait être utilisée à des fins qui vont au-delà de celles qui sont mentionnées dans le cahier des charges. L'appréciation de l'évaluabilité ne doit pas nécessairement être associée à l'organisation d'une évaluation. Ainsi, un test d'évaluabilité appliqué correctement peut par exemple générer des effets importants en termes d'apprentissage et contribuer de cette manière à une meilleure pratique. Plus généralement, le fait d'analyser une intervention sous l'angle de l'évaluabilité peut contribuer de nombreuses manières à améliorer la politique et la pratique dans la coopération au développement.*
   Cette étude recommande dès lors que tous les acteurs, à leur niveau, intègrent d'une façon plus systématique l'évaluabilité et l'appréciation de l'évaluabilité dans leur mode de fonctionnement et considèrent les deux avant tout comme un moyen de rendre la coopération au développement plus performante. L'utilisation de l'évaluabilité ne doit pas se transformer en un levier de contrôle ou de direction bureaucratique (dans le chef du bailleur de fonds, au sein des organisations) mais doit faire partie intégrante des processus de changement et d'apprentissage qui sont déjà en marche dans bon nombre d'organisations. De même, il ne s'agit pas de viser une évaluabilité maximale : l'amélioration de l'évaluabilité doit être une préoccupation permanente, mais doit se situer de manière adéquate dans un contexte spécifique ; il y aura toujours un point critique au-delà duquel le coût de la réalisation d'une meilleure évaluabilité ne contrebalance plus les avantages.

2. *Le suivi et l'évaluation se développent (ou se sont déjà développés) en systèmes bien élaborés sur le plan institutionnel, avec une politique claire et (souvent) des méthodes standardisées. Cette évolution est positive et démontre que l'on a foi dans la valeur de l'évaluation (et du suivi) en tant que composante d'une bonne pratique de développement. Dans ce contexte existe toutefois le risque que la valeur et les bénéfices – entre autres – de l'évaluation (ex ante) ne soient plus remis en question et ne soient plus analysés ex post, ce qui débouche notamment sur des évaluations rituelles ou des évaluations dont la valeur ou l'utilité est limitée, sans que l'on se pose trop de questions à ce sujet.*
   Cette étude recommande par conséquent que lors de chaque évaluation ex ante, une action explicite soit entreprise pour analyser et démontrer les bénéfices potentiels d'une évaluation, plutôt que d'estimer implicitement que ces bénéfices sont toujours présents lors des évaluations, ceci afin de parvenir à une décision fondée quant à la réalisation ou non d'une évaluation. L'introduction d'une appréciation cohérente de l'évaluabilité constitue un instrument important dans ce processus ; elle peut être

réalisée par les acteurs concernés, assistés le cas échéant de l'évaluateur de l'intervention qui a été sollicité.

3. *L'étude a mis en lumière des manquements importants dans la phase préparatoire de nombreuses interventions : les groupes cibles ne sont décrits et délimités que de façon superficielle (en tant que groupes homogènes, sans tenir compte des éléments de différenciation sociale), des acteurs importants de l'intervention ne sont aucunement décrits, ou seulement de façon très sommaire, et de nombreuses analyses sont élaborées pour justifier des choix stratégiques déjà opérés antérieurement (alors que ces choix devraient précisément être basés sur ces analyses). Ces manquements ont d'importantes conséquences pour la qualité du plan d'intervention et pour la suite du cycle d'intervention, ainsi que pour l'évaluabilité.*
   Cette étude recommande une amélioration de la phase préparatoire, en s'attachant plus à la qualité, plutôt que de se limiter à la routine et aux vieilles recettes. Étant donné que ce processus est assez ardu, il est important de mettre en place un processus de changement en vue d'une amélioration progressive, soutenu de différentes manières : via un cadre adapté (avec des incitants) de la DGD, via des bonnes études et évaluations qui peuvent soutenir la formulation (p. ex. à la fin des phases précédentes) et qui se concentrent sur les questions clés et les domaines sur lesquels la connaissance est encore insuffisante, et via une réduction des exigences administratives et des réglementations (liées aux propositions et rapports d'intervention) qui ne contribuent pas à l'efficacité du développement. Cette simplification administrative se conformera de préférence aux principes de la récente 'Note stratégique Résultats de développement' qui stipule que la politique de résultats de la DGD sera axée sur les outcomes plutôt que sur les inputs et outputs.

4. *Il ressort de l'analyse des 40 interventions qu'au niveau opérationnel (inputs, activités, outputs,…), de grands progrès ont été réalisés sur le plan du S&E, ce qui est bénéfique pour l'évaluabilité. Toutefois, pour diverses raisons, on ne consacre que peu d'attention au niveau impact, tandis que le niveau outcome (qui coïncide avec l'objectif spécifique des interventions) est interprété de diverses manières et ne vise pas toujours des changements effectifs au niveau du groupe cible. Ceci est problématique : la coopération au développement implique en effet, par essence, un changement sociétal, or dans bon nombre d'interventions, ce type de changement ne peut être démontré ou ne constitue même pas un objectif explicite.*
   L'étude plaide dès lors (conformément, du reste, avec la Note stratégique Résultats de développement) pour que les niveaux outcome et impact bénéficient d'une attention accrue et soient revalorisés tout au long du cycle d'intervention (plan d'intervention, S&E, …). Ceci implique en premier lieu une définition claire de ces notions de base et de la manière dont elles doivent être appliquées. Plus spécifiquement, l'étude recommande :
   - de veiller, dans les propositions d'intervention, à ce que les outcomes (le niveau de l'objectif spécifique dans le cadre logique) soient définis en tant que *changements* effectifs (dans la politique, les attitudes, la pratique) au niveau des groupes cibles et des institutions locales ; il s'agit des résultats intermédiaires qui – via un lien causal – contribuent à l'impact recherché ;
   - de veiller également, dans les propositions d'intervention, à ce que la TdC entre les effets et impacts (actuels) soit bien élaborée en accordant – par exemple – une attention accrue, d'une part, aux outcomes directs (*intermediate*) auxquels les interventions peuvent, sur la base d'une théorie de changement claire, contribuer de façon démontrable et d'autre part, aux effets à long terme sur le plan sociétal ; une telle TdC peut alors devenir un instrument important pour une évaluation visant à déterminer dans quelle mesure les changements constatés (outcomes et impacts) peuvent être attribués à une intervention (ou sont la conséquence d'autres influences) ;
   - (en lien avec les développements internationaux) définir plus largement l'efficience, de sorte que les outcomes et les impacts soient aussi analysés quant au rapport entre les coûts et les bénéfices.

Cette 'revalorisation' du niveau outcome et impact doit évidemment être prolongée dans la pratique du cycle d'intervention et l'instrumentation qui est actuellement utilisée dans les différentes phases (propositions d'intervention, rapports, système S&E, …).

5. *L'étude a révélé que de nombreuses interventions – en conséquence, bien souvent, de la politique des organisations exécutantes – attachent une importance croissante au développement de la politique et des systèmes S&E et des mécanismes sous-jacents de justification de l'utilisation des moyens reçus. Cette évolution est positive, mais le développement et le fonctionnement de ces systèmes demandent du temps et des moyens assez conséquents de la part de collaborateurs qui, souvent, sont déjà surchargés. Pour cette raison (mais aussi en raison de l'attention limitée accordée au niveau outcome et impact – voir la recommandation 4), il apparaît que la pratique S&E reste souvent bloquée au niveau mise en œuvre : l'attention se porte sur les inputs et outputs relativement faciles à suivre, au détriment du suivi des processus de changement (plus complexes). En corollaire, les systèmes S&E consacrent une attention et des moyens excessifs au suivi, au détriment de l'évaluation. Ce développement est bénéfique pour l'évaluabilité de l'efficience, mais rend plus difficile l'évaluabilité de l'efficacité et surtout de l'impact et de la durabilité.*
   L'étude recommande que les partenaires exécutants de la coopération belge développent, en concertation, une politique, une stratégie et une pratique S&E "évolutives" qui peuvent/doivent commencer par l'élaboration de systèmes et de pratiques au niveau mise en œuvre (telles qu'ils existent déjà actuellement dans bon nombre d'organisations) mais ne doivent pas se limiter à cela. Il est nécessaire, au contraire, que ces systèmes et pratiques continuent à se développer progressivement et visent notamment une bonne évaluabilité de l'efficacité, de l'impact et de la durabilité. Un tel développement (et l'augmentation de l'évaluabilité qui l'accompagne) doit donc idéalement s'opérer de façon graduelle, avec tour à une tour une augmentation des moyens, des instruments, de la capacité et de l'expérience, afin que petit à petit, des fonctions plus complexes puissent être reprises et intégrées. Un bon suivi opérationnel pose ainsi des bases solides pour des évaluations qui peuvent s'orienter vers des questions et des aspects spécifiques qui ne sont pas pris en charge par le suivi.
   Dans le contexte actuel, il est important d'accorder soigneusement cette recommandation avec le screening programmé des ANG et l'harmonisation et la certification des systèmes de suivi et d'évaluation internes des acteurs de la coopération belge (voir aussi la recommandation 7 ci-après).

6. *Les recommandations proposées ci-dessus sont, à différents égards, passablement ardues et difficilement réalisables si l'on tient compte des limitations auxquelles les acteurs de la coopération belge sont confrontés. Ces limitations sont liées tout à la fois à un manque quantitatif de moyens et à un manque d'expérience et d'expertise (qualitatives).*
   C'est pourquoi il est important qu'un cadre soit créé, dans lequel ces changements ambitieux seront non seulement facilités, mais aussi encouragés et valorisés. La DGD joue à cet égard un rôle crucial et pourrait, en accord avec les autres acteurs clés :
   - poursuivre la révision et la simplification de la réglementation, des instruments et des procédures actuels afin qu'ils soient plus axés sur les *effets* (recherchés) de développement et un allègement des obligations en matière de rapports administratifs et financiers ;
   - développer des incitants pour pousser plus loin le développement de la fonction S&E (en particulier la fonction d'évaluation), les acteurs étant ainsi mieux en mesure de (faire) réaliser des évaluations de bonne qualité comportant aussi une analyse de la durabilité et de l'impact ; un important incitant initial peut être, à cet égard, de permettre (notamment dans les nouvelles interventions) que les baselines soient développées au début de l'intervention plutôt qu'au cours de la phase préparatoire ;
   - créer un fonds pour le financement d'études et d'évaluations au niveau effet et impact, dont l'initiative émane – de préférence – de l'ensemble des acteurs

belges du développement. Ce fonds devrait financer des exercices *communs* dans lesquels différentes interventions et différents acteurs seraient impliqués et réaliseraient des études et des évaluations qui dépassent les moyens et les capacités des acteurs individuels et/ou présentent moins d'intérêt direct pour eux. Le SES peut (contribuer à) assurer, dans ce cadre, le contrôle de la qualité sans toutefois être responsable de la gestion de ces évaluations. L'action de ce fonds devrait aussi être dictée par la nécessité, pour la coopération belge, de réaliser plus d'évaluations de type "*public good*".

7. *La certification planifiée des systèmes S&E des acteurs belges du développement est une donnée importante dans le cadre des constatations et des recommandations de cette étude, et nous rappelons à cet égard que l'un des objectifs de cette étude (tels que formulés dans le cahier des charges) est de "… produire des enseignements utiles à l'harmonisation et la certification des systèmes d'évaluation des acteurs …". De même, il est important de signaler le screening programmé (2016) des ANG (dans l'optique de leur accès au cycle de financement suivant), où le système S&E constitue l'un des éléments du screening. Le SES est chargé par le législateur de l'exécution de la certification planifiée, mais n'est pas impliqué dans le screening des ANG.*
   À cet égard, compte tenu des constatations de cette étude, il est recommandé ce qui suit :
   - cette étude a démontré que la qualité des systèmes S&E (en cours de développement et d'exécution) est fortement influencée par une conjonction d'autres éléments de la gestion de l'intervention et du contexte institutionnel et qu'à l'inverse, le système S&E influence à son tour la qualité de ces autres composants de la gestion. En conséquence, il ne semble pas indiqué de dissocier – comme c'est prévu actuellement – la certification de ces systèmes d'une approche plus large et plus intégrée (telle qu'elle est mieux garantie, notamment, dans le screening des ANG, du moins initialement) ;
   - cette étude a aussi montré que la DGD a joué (et peu jouer dans le futur) un rôle important dans l'amélioration qualitative de la gestion de la coopération belge (et indirectement dans l'amélioration de la performance de cette coopération). Les constatations de cette étude (qui, du reste, semble coïncider avec celles d'une évaluation de l'impact des actions des ANG, toujours en cours) devraient avoir pour effet que la DGD conserve des moyens (humains et autres) suffisants pour continuer à remplir ce rôle important. Quant à savoir si une certification formelle est la meilleure approche pour remplir ce rôle, l'avenir le dira. Cette étude a en tout cas permis d'établir que de très nombreux acteurs avaient déjà entamé l'élaboration de systèmes S&E, notamment en vue du screening programmé dont les modalités ont été fixées par le biais d'une concertation entre la DGD et les acteurs concernés. Il a également été constaté que les progrès dans l'élaboration des systèmes S&E au sein des organisations pouvaient varier fortement (entre pays, entre différentes interventions dans un même pays) et que le contexte local (p. ex. dans les pays fragiles) pouvait être déterminant à cet égard. Il s'agit ici, par essence, de processus qualitatifs où le contexte constitue un paramètre important et où la progressivité et le développement de l'*ownership* sont (doivent être) des éléments centraux. Il existe un risque sérieux qu'une démarche de certification qui, inévitablement, doit être exécutée de façon assez formaliste et standardisée et fournit principalement une "photo" de la situation à un moment déterminé (sans lien avec le processus en cours), ait un effet perturbateur par rapport aux processus qui sont actuellement en cours et génère au final des effets contre-productifs du point de vue de l'évaluabilité. Cela ne veut pas dire que la DGD doit laisser entièrement aux acteurs la poursuite du développement des systèmes S&E. La DGD doit pouvoir, si elle le souhaite, avoir accès aux résultats des processus S&E en cours. Elle pourrait pour cela, par exemple, mettre des moyens à disposition pour l'élaboration et l'application (par les acteurs eux-mêmes, ou avec un accompagnement externe) d'un instrument de diagnostic permettant aux

acteurs concernés (et à la DGD) de se faire une idée des points forts et des points faibles de leur système S&E, leur permettant de développer un plan par étapes 'sur mesure' axé sur les résultats en vue d'améliorer leur système et leur pratique S&E, ceci en indiquant de quelle manière et dans quels délais des composantes spécifiques du système et de la pratique S&E seront réajustées et en demandant aux acteurs de faire rapport à la DGD sur l'état d'avancement du plan par étapes.

- dans le prolongement du point précédent et dans l'hypothèse où l'on choisit d'exécuter la certification conformément au plan actuel, il faut éviter que cette certification se transforme en une démarche purement de type 'méta-évaluation' du S&E dans le secteur. Par ailleurs, il est important d'avoir à l'esprit que si l'on procède à la certification des systèmes S&E, il restera de toute façon important, dans le futur, de continuer à effectuer des évaluations externes suffisamment nombreuses et étendues de la qualité du travail des acteurs du développement (en complément aux évaluations initiées par les acteurs eux-mêmes – voir aussi les recommandations opérationnelles ci-après). Pour cela, il y a toutefois d'autres instruments à disposition (ou qui doivent être mis à disposition – voir notamment les recommandations 6 et 11).

### 5.2.2 Recommandations opérationnelles

À partir de l'analyse des facteurs qui influencent l'évaluabilité (voir les chapitres 3 et 4), on peut assez facilement déduire une série de recommandations en matière de 'bonnes pratiques' qui soient bénéfiques non seulement pour l'évaluabilité, mais aussi pour la performance en termes de développement en général. C'est pourquoi nous nous limiterons ici à quelques recommandations opérationnelles importantes qui ont, pour la plupart, des implications dans les différentes phases du cycle d'intervention et/ou sont à nos yeux d'une grande importance opérationnelle. Plusieurs de ces recommandations recouvrent en partie les recommandations stratégiques présentées ci-avant.

8. **Améliorations au plan d'intervention**
   Comme indiqué précédemment (voir la recommandation 3), ces améliorations seront de préférence envisagées comme un processus graduel. Voici à cet égard quelques étapes initiales importantes (pour autant qu'elles soient requises) :
   - une meilleure description (différenciation) et délimitation des groupes cibles directs ;
   - le développement (ou l'amélioration) d'une baseline soigneusement élaborée où (1) les résultats du point précédent sont pris en compte, (2) une définition et un ajustement graduel (vers le "haut") de la théorie de changement interviennent et, par analogie, (3) la formulation des effets (objectif spécifique) et de l'impact (y compris les indicateurs correspondants) recherchés et des hypothèses cruciales est affinée et/ou complétée, (4) des données de base relatives aux indicateurs retenus sont également collectées.

9. **L'ajustement de la politique S&E et sa traduction dans la pratique**
   *Un nombre croissant d'organisations ont élaboré une politique S&E. Dans de nombreux cas, cette politique et surtout son application dans la pratique révèlent toutefois une série de manquements qui peuvent varier d'une organisation à l'autre. Les recommandations qui suivent concernent des manquements qui surviennent assez fréquemment, mais qui ne concernent pas nécessairement tous les intéressés.*
   Dans le développement et la mise en œuvre de la politique S&E, il est important :
   - de développer plus résolument la fonction d'évaluation (au sens strict), en ce compris la relation avec le suivi ;
   - d'accorder plus d'attention à la mise en œuvre de la politique au niveau pays et au niveau intervention, en prenant en compte les spécificités locales (partenaires locaux impliqués dans la réalisation de la politique et de la pratique locales, articulation entre la politique S&E de l'organisation et celle des acteurs locaux) ;

- d'élaborer une approche claire où les efforts et les résultats en matière de S&E sont utilisés pour une analyse, un apprentissage et un changement effectifs.

10. **Développement de la fonction S&E au niveau intervention**
    Outre la nécessite d'une attention accrue pour les 'niveaux supérieurs' dans la chaîne moyens-fin et pour l'évaluation (voir les recommandations 4 et 5), il est important :
    - que l'organisation et le fonctionnement des systèmes S&E soient bien planifiés ; ceci n'implique pas nécessairement que ces systèmes soient déjà élaborés pendant la formulation de l'intervention, mais bien que leur organisation, leur développement et leur fonctionnement soient bien définis et planifiés, avec une attention pour les moyens humains et financiers nécessaires, le renforcement des capacités, etc. ;
    - que les différentes composantes du S&E soient réellement réunies en un *système* cohérent qui englobe tout le champ S&E, évite les chevauchements entre les différentes composantes sur le plan du contenu et implique une bonne définition des moyens nécessaires (sur le plan quantitatif et qualitatif) et de la répartition des tâches (rôle des différents acteurs) ;
    - que le rôle des différents acteurs dans les systèmes S&E soient optimisés, ceci en évitant que le S&E devienne le monopole des partenaires responsables de la mise en œuvre et en déterminant, en partant plutôt du principe de subsidiarité, qui peut contribuer au S&E, où et de quelle manière ;
    - qu'une attention suffisante soit consacrée au feedback sur les résultats du suivi envers toutes les parties concernées, ceci afin d'alimenter constamment la sensibilisation à l'importance de la (bonne) collecte de données et de contribuer à la pérennisation du système ;
    - de vérifier comment le système S&E de l'intervention s'articule par rapport aux systèmes S&E locaux et nationaux et, si nécessaire et si ceci est pertinent, quel rôle l'intervention peut jouer dans la mise en place, l'amélioration et l'ancrage de ces systèmes locaux et nationaux.

11. **Améliorer et exécuter les évaluations en fonction d'une attention permanente pour leur utilisation finale**
    *Cette recommandation s'appuie – entre autres – sur la recommandation 2 et part du postulat selon lequel la qualité d'une évaluation dépend en premier lieu de son utilisation effective, déterminée principalement par sa valeur et son utilité. À cet effet, une réflexion en profondeur sur les bénéfices futurs de l'évaluation (recommandation 2) est une condition nécessaire mais insuffisante.*
    Afin d'optimiser l'utilisation finale des évaluations, il est important :
    - conformément à la recommandation 2, de prévoir dans chaque évaluation une appréciation de l'évaluabilité qui analyse et démontre, entre autres, les bénéfices potentiels d'une évaluation planifiée ;
    - d'inscrire la planification et l'exécution d'une évaluation dans une "approche de portefeuille', dans laquelle il ne s'agit pas d'évaluer toutes les interventions dans tous leurs aspects, mais d'opérer des choix stratégiques en fonction des moyens disponibles, des exigences sur le plan de la redevabilité, des facteurs contextuels, etc. Il peut aussi être intéressant, par exemple, de répartir les interventions en différents groupes en fonction des connaissances dont on dispose déjà concernant le fonctionnement et l'impact de l'intervention :
        - des interventions dont des évaluations et études antérieures ont démontré à maintes reprises qu'elles ont un impact (souvent des interventions avec une TdC moins complexe) ne doivent plus à chaque fois être évaluées quant à leur impact. Pour ce groupe d'interventions, il n'est donc pas indispensable qu'un counterfactual soit spécifié et il suffit d'évaluer l'efficacité et l'efficience ;
        - des interventions dont on sait qu'elles sont correctement mises en œuvre et que l'on veut reproduire ou étendre substantiellement, mais

dont on ne sait pas suffisamment si elles génèrent un impact, seront de préférence soumises à une évaluation d'impact approfondie. Il est nécessaire que ces interventions puissent aussi être effectivement évaluées quant à leur impact et qu'il en soit tenu compte dès la conception de l'intervention. Ceci implique que les différents aspects de l'évaluabilité de l'impact qui ont été analysés tout au long de l'étude fassent l'objet d'une attention réelle (notamment TdC jusqu'au niveau de l'impact, identification des éléments cruciaux jusqu'au niveau de l'impact, traduction de la TdC en un système S&E jusqu'au niveau de l'impact, spécification d'un counterfactual, bonne baseline, etc.). En particulier, la présence de données pour un counterfactual[64] et concernant les hypothèses externes est importante avant de procéder à une évaluation d'impact[65] ;
    - pour des interventions dont ne sait pas suffisamment si le mode de mise en œuvre qui a été choisi est le meilleur, il convient d'essayer différents modes au cours de la mise en œuvre et de s'engager dans l'évaluation qui compare le fonctionnement et l'efficacité de ces différents modes (voir aussi, à cet égard, l'idée de *structured experiential learning*, note de bas de page 28).
- d'impliquer les utilisateurs finaux de l'évaluation (acteurs impliqués dans l'intervention et autres) dès la préparation de l'évaluation (par exemple en prenant en compte leurs intérêts et leurs attentes) ;
- de rédiger des termes de référence avec des questions clés principales bien délimitées et soigneusement étudiées et qui permettent aux évaluateurs (voire les y incitent) de mener des recherches en dehors du cadre strict des objectifs et des questions principales de l'évaluation.

[64] Ceci ne requiert pas nécessairement l'exécution d'un RCT (Random Controlled Trial). Au chapitre 2, différentes possibilités ont été évoquées pour l'identification d'un counterfactual.

[65] S'il est encore possible de remédier, pendant une évaluation de l'impact, à l'absence d'une TdC explicite jusqu'au niveau de l'impact (dans la mesure où les responsables des interventions ont souvent une TdC implicite sans que celle-ci soit mise sur papier), l'absence, dans la pratique, d'un système S&E qui collecte des données valides pour le groupe cible/la situation de l'intervention-groupe/situation et pour un counterfactual est plus difficile à compenser.

# Annexes

1. Cahier des charges
2. Description de la méthodologie de recherche et de l'approche
3. Cadre d'étude
4. Liste des 40 interventions analysées
5. Liste des principaux documents consultés
6. Liste des principales personnes contactées
7. Commentaires sur les analyses statistiques et l'analyse des items les plus forts et les plus faibles
8. Aperçu détaillé des scores